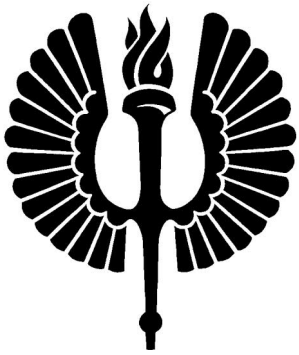


Monivertailun ongelmista

Ida Isaksson

Pro gradu -tutkielma
Joulukuu 2018

MATEMATIIKAN JA TILASTOTIETEEN LAITOS
TURUN YLIOPISTO

TURUN YLIOPISTO
Matematiikan ja tilastotieteen laitos

ISAKSSON, Ida: Monivertailun ongelmista
Pro gradu -tutkielma, 34 s., 10 liites.
Tilastotiede
Joulukuu 2018

Erityisesti geeni- ja mikrojonodatat ovat vauhdittaneet monivertailutestaamisen kehitystä. Teknologian nopea kehittyminen on mahdollistanut suurten aineistojen käsittelyn, mikä onkin nostanut esiin tyypillisiä monivertailun ongelmia. Tutkijan kannalta yksi keskeisimmistä ongelmista onkin hylkäysvirheiden käsittely.

Yksittäisellä testillä on ennalta määritelty merkitsevyystaso, joka kertoo suurimman sallitun todennäköisyyden tehdä hylkäysvirhe, eli todennäköisyyden hylätä virheellisesti tosi nollahypoteesi. Mitä enemmän testejä suoritetaan samanaikaisesti, sitä suuremmaksi kasvaa todennäköisyys, että yksi tai useampi nollahypoteesi hylätään virheellisesti. FDR-menetelmä, eli hylkäysvirheasteen rajoittamiseen pyrkivä menetelmä, on noussut yhdeksi suosituimmista tavoista käsitellä hylkäysvirheitä.

FDR-menetelmä pyrkii rajoittamaan tosien nollahypoteesien odotusarvoista osuutta kaikkien hylättyjen nollahypoteesien joukossa. Kyseisen menetelmän avulla saavutetaan enemmän merkitseviä löydöksiä kuin esimerkiksi Bonferroni-korjauksella, kuitenkin niin, että hylkäysvirheaste pysyy halutulla tasolla.

Perinteisemmät monivertailumenetelmät, esimerkiksi Bonferroni-korjaus, keskittyvät hallitsemaan ainoastaan tyypin I virheen todennäköisyyttä. Suurten aineistojen tapauksessa nämä menetelmät ovat kuitenkin usein todella konservatiivisia.

Asiasanat: hypoteesien testaaminen, monivertailu, merkitsevyystaso, FDR-menetelmä

Sisältö

1 Johdanto

Tilastollisen laskennan tehokkuus on kehittynyt huimasti 1900-luvun loppupuolella. Tietokoneiden laskentatehon kehitys on antanut uuden näkökulman tilastotieteen harjoittamiseen. Erityisesti geenitutkimuksien yleistyminen on tutustuttanut tilastolliset tutkijat ns. suurten datojen analysoimiseen [1]. Mikrosirujen hyödyntäminen biolääketieteellisessä tutkimuksessa tarjoaa pääsyn tuhansien geenien tutkimiseen samanaikaisesti. Geenien vertailemista varten täytyy muodostaa tuhansia tilastollisia testejä, minkä vuoksi on syntynyt tarve suuren datan analysoimisen menetelmille.

Monivertailussa täytyy ottaa huomioon vastahypoteesin väärän hyväksymisen todennäköisyys, joka kasvaa sitä mukaan, mitä useampia testejä suoritetaan samanaikaisesti. Tavallisessa tilastollisessa tutkimuksessa tutkittavan ilmiön tai asian perusteella muodostetaan hypoteesit, joita testataan ja tuloksesta riippuen hyväksytään tai hylätään. Monivertailussa hypoteesipareja muodostetaan tapauksesta riippuen jopa kymmeniätuhansia, minkä vuoksi myös testejä täytyy muodostaa kymmeniätuhansia.

Yksittäiselle tilastolliselle testille määritellään merkitsevyystaso, minkä perusteella nollahypoteesi hylätään tai hyväksytään. Jos yhden testin merkitsevyystasoksi asetetaan 0.05, tällöin todennäköisyys epätoden nollahypoteesin hyväksymiselle on 5%. Monivertailussa testejä suoritetaan useampia kuin yksi, minkä vuoksi todellisuudessa todennäköisyys vähintään yhden tai useamman nollahypoteesin virheelliselle hyväksymiselle on suurempi kuin 5%. Mitä enemmän testejä suoritetaan, sitä suuremmaksi todennäköisyys kasvaa. Monivertailua varten on kehitetty monenlaisia tilastollisia menetelmiä, jotka kontrolloivat virheellisen hylkäämisen todennäköisyyttä. Suuren datan tapauksessa nämä menetelmät ovat kuitenkin liian konservatiivisia, minkä vuoksi suosiotaan on kasvattanut menetelmä, joka kontrolloi hylkäysvirheastetta (engl. *False-Discovery Rate*). Hylkäysvirheaste on odotettu virheellisten löydösten eli tosien nollahypoteesien osuus kaikkien hylättyjen nollahypoteesien joukossa.

2 Hypoteesien testauksesta

Klassillisessa tilastollisessa testaamisessa tehdään päätelmiä ilmiöistä hypoteesien ja havaintojen eli otoksen perusteella. Ensin muodostetaan nolla- ja vastahypoteesit. Nollahypoteesi on jokin koko kohdeväestöä koskeva väite, joka yritetään kumota otoksien ja tilastollisen testin avulla. Vastahypoteesi on vastaväite nollahypoteesille. Tutkittaessa kahden ryhmän, esimerkiksi miesten ja naisten painojen eroa, nollahypoteesiksi asetetaan

$H_0 = \{$naisten ja miesten populaatioiden keskimääräinen paino on yhtä suuri$\}$.

Vastahypoteesi voidaan muodostaa vastaavasti

$H_1 = \{$naisten ja miesten populaatioiden keskimääräinen paino on eri suuri$\}$.

Kun hypoteesiparit on muodostettu, valitaan tutkittavaan ilmiöön soveltuva testi. Tämä tapahtuu usein tutkimalla ryhmien välistä riippuvuutta sekä tarkastelemalla aineiston kuvaajia. Seuraavaksi valitaan testin merkitsevyystaso α, eli suurin sallittu todennäköisyys hylätä nollahypoteesi, joka on tosi. Testin valitsemisen jälkeen voidaan laskea testisuureen jakauma sekä testisuureen kriittinen arvo, johon verrataan aineistosta laskettua testisuureen arvoa. Vertaaminen tapahtuu hylkäysalueen määrittämisellä. Jos testisuureen havaittu arvo osuu hylkäysalueelle, nollahypoteesi hylätään. Hylkäysalue määritellään testisuureen kriittisen arvon sekä merkitsevyystason α avulla.

2.1 Testisuure

Tilastollista testiä tehtäessä testisuureen T määrittäminen on suuressa roolissa, sillä kuten aikaisemmin mainittiin, testin lopputulos määräytyy testisuureen havaitun arvon vertaamiseen testisuureen nollahypoteesin mukaiseen jakaumaan. Testisuureen arvoa käytetään myös p-arvon määrittämiseen, minkä avulla arvioidaan tilastollisen testin merkitsevyys.

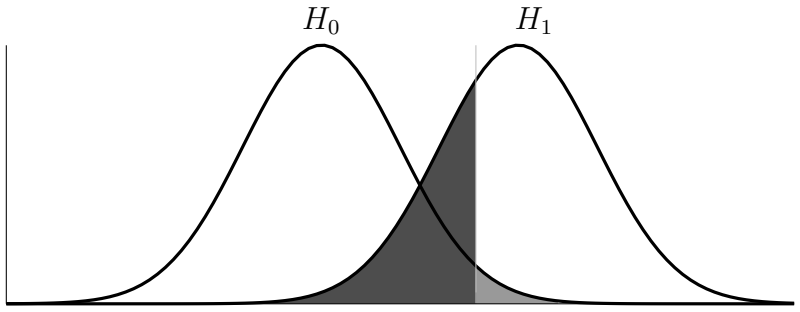
Cox & Hinkley esittävät kirjassaan [11] testisuureen määritelmän seuraavasti. Olkoon $t = t(x)$ havaintojen muodostama funktio ja $T = t(X)$ sitä vastaava satunnaismuuttuja. T on testisuure, jos seuraavat ehdot täyttyvät:

- Testisuureen T jakauma tunnetaan nollahypoteesin H_0 alaisuudessa
- Mitä suurempi havaittu arvo t on, sitä vahvempi todiste tarvitaan nollahypoteesin H_0 hyväksymiseksi. Toisin sanoen mitä suurempi havaittu testisuureen arvo on t, sitä ilmeisimmin nollahypoteesi ei ole voimassa.

2.2 Tyypin I ja II virheet

Tilastollinen testaaminen perustuu todennäköisyyksien laskemiseen, ja usein ilmiötä/tutkittavaa tapahtumaa tarkasteellaan kerätyn otoksen kautta. Edellä mainitun vuoksi testaamisen yhteydessä tapahtuu kahdenlaisia virheitä.

- Tyypin I virhe, eli väärä positiivinen, tapahtuu kun hylätään nollahypoteesi, joka on tosi. Tyypin I virhettä kutsutaan myös hylkäysvirheeksi. Nollahypoteesi hylätään, vaikka se olisi tosi. Kuvassa 1 vaalean harmaa alue kuvaa tyypin I virheen todennäköisyyttä.
- Tyypin II virhe, eli väärä negatiivinen tapahtuu taas, kun hyväksytään nollahypoteesi, joka ei ole tosi. Tyypin II virhettä kutsutaan myös hyväksymisvirheeksi. Kuvassa 1 tumman harmaa alue kuvaa tyypin II virheen todennäköisyyttä.



Kuva 1: Yksittäisen tilastollisen testin hypoteesien mukaiset jakaumat.

Ihanteellisessa tapauksessa molempien virheiden esiintymistä pyritään vähentämään samanaikaisesti. Tämä ei ole kuitenkaan mahdollista, minkä vuoksi virheiden välillä joudutaan tekemään valinta. Yleensä tämä tapahtuu asettamalla tyypin I virheelle tietty todennäköisyys, joka on hyväksyttävää. Yleisimmin tyypin I virheen sallituksi todennäköisyydeksi asetetaan 5%.

Tyypin I virheen todennäköisyyden kontrolloimiseen tarvitaan testisuureen nollahypoteesin mukaista jakaumaa, josta voidaan johtaa hypoteesien hylkäämisalueet.

2.3 Testin p-arvo ja merkitsevyystaso

Olkoon $t = t_{hav} = t(x)$ aineistosta laskettu havaittu testisuure. Määritellään kolmelle eri testille p-arvot:

$$p = P(T \geq t_{hav}|H_0), \text{ oikeanpuoleinen p-arvo, yksisuuntainen testi} \quad (1)$$
$$p = P(T \leq t_{hav}|H_0), \text{ vasemmanpuoleinen p-arvo, yksisuuntainen testi} \quad (2)$$
$$p = 2 * min\{P(T \leq t_{hav}|H_0), P(T \geq t_{hav}|H_0)\}, \text{ kaksisuuntainen testi.} \quad (3)$$

P-arvon voidaan sanoa olevan nollahypoteesin sekä aineiston yhdenmukaisuuden mitta. Symmetrisen kaksisuuntaisen testin tuottama p-arvo kertoo todennäköisyyden, jolla nollahypoteesin ollessa voimassa, havaitaan koetta toistettaessa itseisarvoltaan vähintään yhtä suuri tai suurempi testisuureen arvo kuin havaittu arvo. [11]

Lähellä lukua yksi olevat p-arvot puoltavat nollahypoteesia. Pienet p-arvot puoltavat vastahypoteesia, sillä nollahypoteesin vallitessa on epätodennäköistä havaita testin suunnasta riippuen, vähintään yhtä suuri (tai yhtä pieni) testisuureen arvo, kuin havaittu arvo. [12]

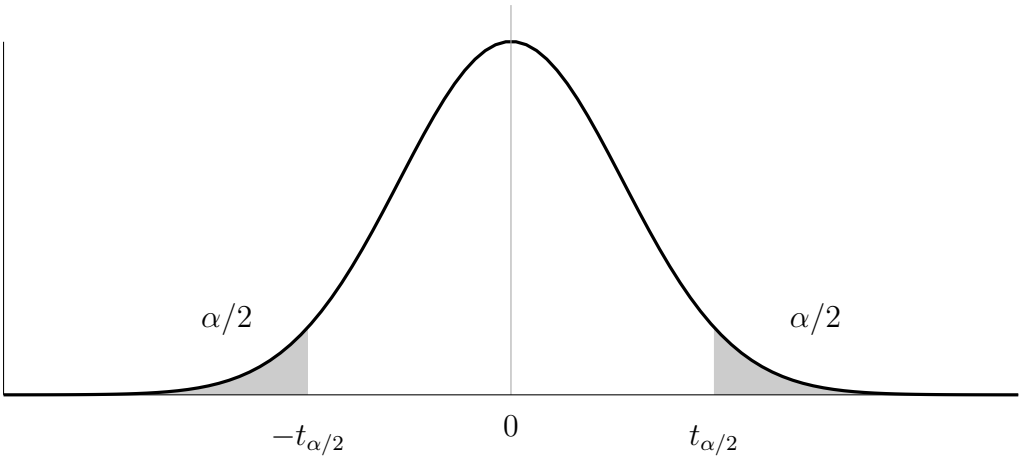
P-arvoa voidaankin verrata suoraan ennalta määriteltyyn merkisevyystasoon. Merkitsevyystaso on todennäköisyys hylätä nollahypoteesi, joka on tosi. Yksittäiselle testille merkitsevyystaso määritellään

$$P(H_0 \text{ hylätään}|H_0 \text{ on tosi}) = \alpha. \quad (4)$$

Huomioitavaa on, että merkitsevyystaso α on ennalta määritelty suurin sallittu todennäköisyys hylätä tosi nollahypoteesi. P-arvo taas on aineistosta laskettu suure, jota verrataan merkitsevyystasoon α. Jos saatu p-arvo on pienempi kuin ennalta määritelty merkitsevyystaso α, nollahypoteesi hylätään. Jos p-arvo on suurempi kuin α, nollahypoteesi jää voimaan. Huomioitavaa on, että vaikka p-arvo olisi kuinka lähellä lukua yksi, niin voidaan ainoastaan todeta, että nollahypoteesiä ei onnistuttu hylkäämään havaitun datan perusteella. On silti mahdollista, että nollahypoteesi ei pidä paikkaansa todellisuudessa. [12] Nollahypoteesin vallitessa p-arvo on satunnaismuuttujan P havaittu arvo, missä $P \backsim \mathcal{U}(0, 1)$.

Havainnoillistetaan hypoteesien testaamista tarkastelemalla yksinkertaista hypoteesiparia $H_0 : \mu = \mu_0$ ja $H_1 : \mu \neq \mu_0$. Olkoon otoskoko n, havaintojen keskiarvo $\bar{X}$, ja oletetaan, että havainnot ovat normaalijakautuneita sekä toisistaan riippumattomia. Tarkoituksena on selvittää, onko jakauman odotusarvo 0. Näiden oletusten perusteella kaksisuuntainen t-testi on sopiva näiden hypoteesien testaamiseen. Yhden otoksen t-testisuure on muotoa

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}, \quad (5)$$



Kuva 2: Kaksisuuntaisen yhden otoksen t-testin hylkäysalueet.

missä s on otoksen keskihajonta. Testi suoritetaan merkitsevyystasolla $\alpha = 0.05$. T-testin oletusten mukaan testisuure $t \backsim t(n-1)$, kun nollahypoteesi on tosi. Olkoon otoskoko 30 ja oletetaan, että testisuuren arvoksi on saatu $t = 3.234$. Testi on kaksisuuntainen t-testi, joten hylkäämisaluetta varten määritellään kriittiset rajat $-t_{\alpha/2}$ ja $t_{\alpha/2}$, jotka toteuttavat ehdon

$$P(-t_{\alpha/2} \leq t \leq t_{\alpha/2}) = 1 - \alpha, \tag{6}$$

missä α on ennaltamääritelty merkitsevyystaso. Kuvassa 2 on havainnoillistettu kaksisuuntaisen t-testin hylkäämisalueet.

Esimerkissä nollahypoteesin mukaisesti testisuure noudattaa Studentin jakaumaa vapausastein $n - 1 = 29$, jolloin t-jakauman kvantiilien $-t_{0.025}$ ja $t_{0.025}$ arvot ovat -2.045 ja 2.045. Nyt hylkäysalue määritellään

$$(\infty, -2.045) \cup (2.045, \infty). \tag{7}$$

Koska havaittu testisuureen arvo osuu hylkäämisalueelle $3.234 > 2.045$, nollahypoteesi hylätään, ja vastahypoteesi astuu voimaan. Studentin t-jakauma on symmetrinen, joten p-arvo voidaan yhtälön (3) mukaan määritellä

$$p = 2 * P(T > |t|) = 2 * P(T > |3.234|) = 2 * 0.004 = 0.008. \tag{8}$$

2.4 Testin voimakkuus

Tyypin II virhe, eli hyväksymisvirhe tapahtuu, kun nollahypoteesi hyväksytään, vaikka se ei ole tosi. Hyväksymisvirhe voidaan määritellä:

$$P(H_0 \text{ hyväksytään} | H_0 \text{ ei ole tosi}) = \beta. \tag{9}$$

Testin voimakkuus on hyväksymisvirheen komplementti $1-\beta$. J. Kalbfleischin mukaan [12] testin voimakkuus K_α voidaan määritellä

$$\begin{aligned} K_\alpha &= P(p \leq \alpha | H_1 \text{ on tosi}) && (10)\\ &= P(H_0 \text{ hylätään} | H_0 \text{ ei ole tosi}) && (11)\\ &= 1 - \beta. && (12) \end{aligned}$$

Jos testin voimakkuus on suuri, eli K_α on lähellä yhtä, testin sanotaan olevan voimakas, sillä tällöin hyväksymisvirhe β on pieni.

Aikaisemmin esitetty testisuureen valinta on kriittinen vaihe tilastollisessa testaamisessa, minkä vuoksi testisuureita voidaankin verrata keskenään testin voimakkuuden avulla. Mitä suurempi testin voimakkuus saavutetaan, sitä parempi kyseinen testisuure on. Testin voimakkuutta käytetään myös minimiotoskoon valitsemisessa; valitaan haluttu voimakkuus ja lasketaan mikä on pienin otoskoko n, joka saavuttaa kyseisen voimakkuuden.

3 Monivertailu

Kun samanaikaisesti suoritettavia testejä on useampia kuin yksi, puhutaan monivertailusta. Monivertailun tulokset voidaan raportoida samoin termein kuin yhden hypoteesin tapauksessa: laskemalla testin tuottama p-arvo, määräämällä hylkäämisalueet testisuureelle sekä laskemalle luottamusalueet kiinnostuksen kohteena olevalle parametrille. [2]. Monivertailulle on tyypillistä, että testejä suoritetaan samanaikaisesti satoja tai jopa tuhansia. Kun testejä suoritetaan useampia samanaikaisesti, todennäköisyys tehdä vähintään yksi virheellinen päätös kasvaa. Tämän vuoksi on tyypin I virheen kontrolloiminen on erittäin tärkeää. Monivertailussa kunkin hypoteesiparin jakaumat voivat olla erimuotoisia, ja koska kynnysarvo α on kuitenkin kiinnitetty, tyypin I virheen esiintyminen voi vaihdella suuresti hypoteesipareista riippuen. Jos sallituksi tyypin I virheen tasoksi asetetaan 5%, on 95% todennäköisyys, että tutkimuksessa hyväksytään tosi nollahypoteesi. Jos suoritetaankin samanaikaisesti viisi testiä tilanteessa, jossa kaikki viisi nollahypoteesia ovat tosia, tällöin todennäköisyys, että kaikki viisi nollahypoteesit hyväksytään, on enää $0.95^5 = 0.77$, jolloin tyypin I virheen todennäköisyys on 23%. Suurten aineistojen testaamisessa on tyypillistä, että nollahypoteesia noudattavien havaintojen määrä on todella suuri. Tämän vuoksi on erittäin tärkeää pyrkiä pitämään tyypin I virheen esiintyvyys hallitulla tasolla silloinkin, kun testejä tehdään suuri määrä.

Käytetään monivertailun havainnoillistamista varten aineistoa [10], joka löytyy R:stä paketista sda. Aineistoon on kerätty $n = 102$ miestä, joista $n_2 = 52$ sairastaa eturauhassyöpää ja $n_1 = 50$ miestä kuuluu terveeseen kontrolliryhmään. Jokaiselta subjektilta on kerätty havainto $N = 6033$ geenistä, jolloin aineisto koostuu havainnoista x_{ij}, jossa

$$x_{ij} = \text{geenin } i \text{ aktiivisuus miehellä } j. \tag{13}$$

Jos geeni i ei ole aktiivinen, sitä kutsutaan nollageeniksi. Tavallisimmin monivertailu suoritetaan käyttäen varianssianalyysia (ANOVA) tai t-testiä. T-testi soveltuu kahden otoksen keskiarvojen vertailuun, kun taas varianssianalyysia käytetään useamman otoksen keskiarvojen vertailuun. Syöpäaineisto koostuu kahdesta eri ryhmästä, joten tässä tapauksessa on sopivaa käyttää kahden otoksen t-testiä.

Vertailtaessa syöpäpotilaiden ja kontrolliryhmän geenin i aktiivisuutta, eli pyritään selvittämään onko geeni i nollageeni vai ei, on testisuure muotoa:

$$t_i = \frac{\bar{x_i}(2) - \bar{x_i}(1)}{s_i} \tag{14}$$

missä $\bar{x}_i(1)$ ja $\bar{x}_i(2)$ ovat kontrolliryhmän ja syöpäpotilaiden geenin i keskimääräinen aktiivisuus, ja s_i^2 on geenin i otosvarianssi:

$$s_i^2 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x}_i(1))^2 + \sum_{i=1}^{n_2}(x_i - \bar{x}_i(2))^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \tag{15}$$

Monivertailun tapauksissa on usein oletuksena, että vain muutama havainto poikkeaa nollahypoteesista. Seuraavat oletukset pohjatuvat Efronin [8] määritelmiin. Eturauhassyöpätutkimuksessa taustaoletuksena on, että suurin osa geeneistä on nollageenejä, eli syöpäpotilaiden ja kontrolliryhmän geenit eivät poikkea toisistaan. Oletetaan, että $x_{ij} \backsim \mathcal{N}(\mu_i, 1)$. Olkoon geeni i nollageeni, jos syöpäpotilaiden ja kontrolliryhmän geenien aktiivisuustasot eivät poikkea toisistaan. Määritellään nollahypoteesi i:

H_{0i} = geeni i on nollageeni, jos odotusarvo μ_i ei riipu ryhmästä 1 tai 2.

Nollahypoteesin ollessa voimassa, syöpäsoluja kantavien miesten geenit noudattavat samaa normaalijakaumaa kuin kontrolliryhmän miesten geenit. Tällöin testisuure t_i noudattaa Studentin standardoitua t-jakaumaa 100:lla vapausasteella, t_{100}. Tulevia analyyseja varten on käytännöllisempää laskea muunnos

$$z_i = \Phi^{-1}(F_{100}(t_i)), \tag{16}$$

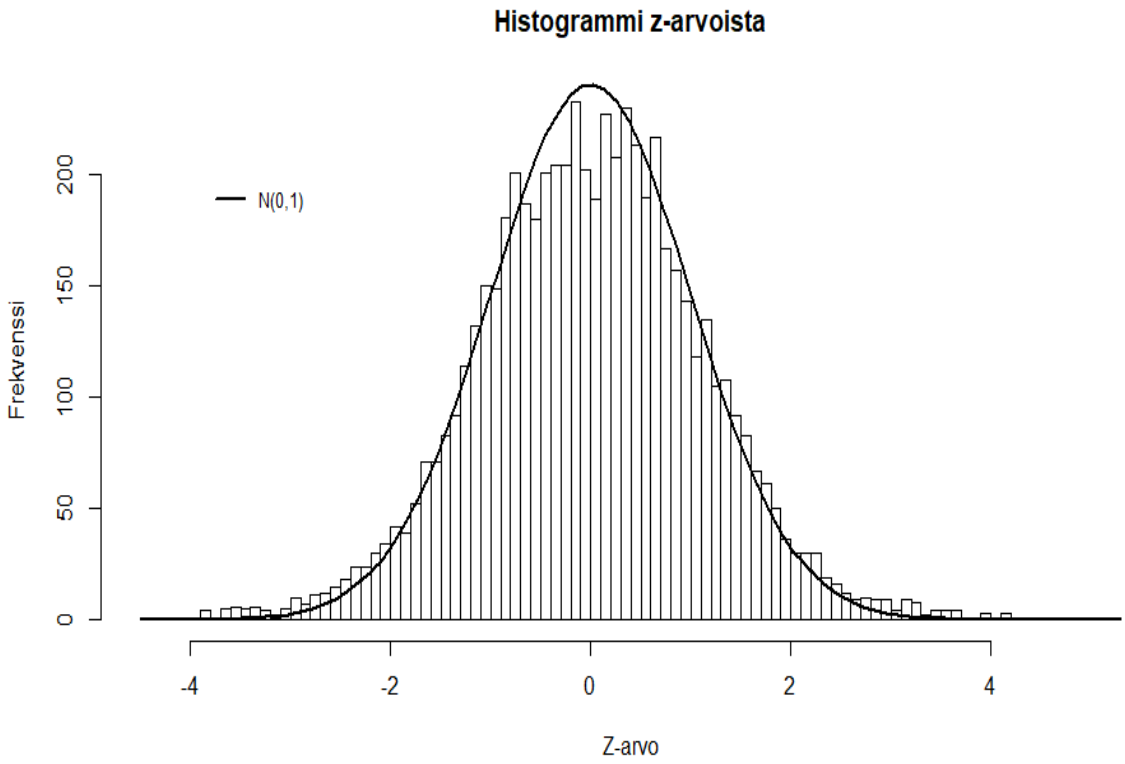
missä F_{100} on t_{100} jakauman kumulatiivinen jakaumafunktio ja Φ^{-1} on standardoidun kumulatiivisen normaalijakauman käänteisfunktio. Nollahypoteesin vallitessa z-arvot noudattavat normaalijakaumaa

$$H_{0i} : z_i \backsim \mathcal{N}(0, 1). \tag{17}$$

Ei-nollageeneiksi kutsutaan niitä geenejä, joiden aktiivisuus on syöpä- ja kontrolliryhmällä erilaista.

Kuvan 3 käyrä esittää normaalijakauman $\mathcal{N}(0, 1)$ tiheysfunktiota. Kuvasta 3 nähdään kuitenkin, että kaikki geenit eivät mukaile normaalijakauman mukaista käyrää. Tavallisimmin tilastollisessa testaamisessa asetetaan merkitsevyystasoksi $\alpha = 0.05$. Tyypin I virheen määritelmän mukaan tällöin suurin sallittu todennäköisyys hylätä tosi nollahypoteesi on 5%. Syöpäaineiston tapauksessa nollahypoteesin vallitessa, z-arvot noudattavat standardoitua normaalijakaumaa, jolloin merkitseväksi valitaan kaksisuuntaisen t-testin tapauksessa ne z-arvot, joille $|z| > 1.96$. Esimerkkiaineiston tapauksessa on 302 z-arvoa, jotka ylittävät edellellä mainitun rajan.

Edellä esitetyn t-testin suorittamiseen käytettiin merkitsevyystasoa $\alpha = 0.05$. Suorittaessa $N = 6033$ testiä samanaikaisesti, hylkäysvirheiden lukumäärän odostusarvo on $6033 * 0.05 = 302$.



Kuva 3: Histogrammi N=6033 z-arvosta. Jos jokainen geeni olisi nollageeni, histogrammi mukailisi käyrää.

3.1 Yhdistetty merkitsevyystaso (FWER)

Monivertailun tapauksessa ongelmana on löytää tarpeeksi tehokas tapa havaita nollahypoteesista poikkeavat havainnot. Monivertailutestit keskittyvät yleensä tyypin I virheen kontrolloimiseen. Yhdistetty merkitsevyystaso (*engl. Family-wise error rate, FWER*) kontrolloi tyypin I virheen esiintymistodennäköisyyttä.

Yhdistetty merkitsevyystaso

$$FWER = P(\text{Hylätään ainakin yksi tosi } H_{0i}) \tag{18}$$

FWER on todennäköisyys, että suoritettaessa N samanaikaista testiä, tehdään vähintään yksi virheellinen hylkäyspäätös.

Oletetaan, että joukkoon I_0 kuuluu N_0 kappaletta tosia nollahypoteeseja H_0. Tällöin yhdistetty merkisevyystaso on

$$FWER = P(\bigcup_{I_0}(p_i \leq \frac{\alpha}{N})) \leq \sum_{I_0} P(p_i \leq \frac{\alpha}{N}) = N_0 \frac{\alpha}{N} \leq \alpha. \tag{19}$$

FWER:n todennäköisyys lasketaan siis kaikkien hypoteesien joukossa. Huomioitavaa on myös, että edellä esitetty lauseke on saavutettu Boolen epäyhtälön avulla [8].

3.1.1 Bonferroni-korjaus

Yleisin FWER:n kontrollointimenetelmistä lienee Bonferroni-korjaus, jonka ideana on kasvattaa yksittäisen testin hylkäämisrajaa [1]. Olkoon samanaikaisesti testattavien hypoteesien lukumäärä N. Bonferroni-korjauksen mukaan i:nnes nollahypoteesi H_{0i} hylätään, jos se saavuttaa yksittäisen merkitsevyystason α/N. Tällöin kaavan (19) perusteella Bonferroni-korjaus kontrolloi siis FWER:ää, riipumatta siitä, kuinka moni nollahypoteeseista H_{0i} on tosi [8]. Mitä suurempi nollahypoteesien määrä N on, sitä konservatiivisemmaksi Bonferroni-korjaus muuttuu.

Klassillinen tilastollinen testaaminen ilmaistaan yleensä p-arvojen ja merkitsevyystasojen avulla. Kun yksisuuntaisen testin kriittinen alue on $z \geq z_0$, p-arvo määritellään:

$$p(z) = 1 - F_0(z). \tag{20}$$

Vastaavasti, kun kriittien alue on $z \leq z_0$, p-arvo määritellään:

$$p(z) = F_0(z). \tag{21}$$

Edellä esitetyissä lausekkeissa $F_0(z)$ on testisuureen z kertymäfunktio. Mitä suurempia arvoja z saa, sitä pienempi p-arvo havaitaan. Nollahypoteesi hylätään, jos saatu p-arvo on pienempi kuin ennaltamääritelty merkitsevyystaso α. Bonferroni-korjattujen merkitsevien valittujen p-arvojen joukko määritellään:

$$\mathcal{R}_i(\alpha) = \{p_i \leq \frac{\alpha}{N}\}. \tag{22}$$

Syöpäaineiston tapauksessa $N = 6033$ ja $\alpha = 0.05$, jolloin Bonferroni-korjauksen mukaan hylätään ainoastaan geenit, joille $p_i \leq 0.0000083$, mitä vastaava kriittinen arvo yksisuuntaisen testin tapauksessa on $z = -4.31$. Edellä laskettu z-arvo perustuu yhtälöön $P(Z \leq z) = 0.0000083$. Aineistosta löytyy ainoastaan 2 z-arvoa, jotka toteuttavat ehdon $z_i \leq -4.31$, jotka siis valittaisiin ei-nollageeneiksi Bonferroni-korjauksen perusteella.

4 Hylkäysvirheaste (FDR)

Edellä esitelty FWER-menetelmä kontrolloi todennäköisyyttä tehdä vähintään yksi virheellinen nollahypoteesin hylkäys suoritettaessa N samanaikaista testiä. FWER-menetelmän on kuitenkin todettu olevan liian varovainen hylkäämään nollahypoteeseja, kun testien määrä on suuri ($N > 20$). [1] Monivertailutilanteissa on tyypillistä, että tosien nollahypoteesien osuus on lähellä lukua yksi. Tämän vuoksi on tärkeää pyrkiä hallitsemaan tosien nollahypoteesien osuutta hylättyjen hypoteesien joukossa sen sijaan, että oltaisiin kiinnostuneita hallitsemaan todennäköisyyttä tehdä tyypin I virhe.

Kun tilastollisia testejä tehdään samanaikaisesti satoja tai jopa tuhansia, Benjaminin & Hochbergin [3] kehittämä FDR-menetelmä (*engl. False Discovery Rate*) on kasvattanut suosiotaan tutkijoiden keskuudessa. FDR-menetelmän avulla hylkäysvirheastetta eli virheellisten löydösten odotettavaa osuutta voidaan estimoida. Tästä eteenpäin hylkäysvirheastetta merkitään lyhenteellä FDR.

Olkoon samanaikaisesti testattavien hypoteesien lukumäärä N, joista tosien nollahypoteesien lukumäärä on N_0. Taulukosta (1) nähdään, että hylättyjen nollahypoteesien kokonaislukumäärä on R, joka on havaittavissa oleva satunnaismuuttuja. Suureet $N_0 - a$, a, $N_1 - b$ ja b ovat sitä vastoin satunnaismuuttujia, joita ei havaita. Taulukon (1) merkintöjä käyttäen, edellisessä kappaleessa esitelty FWER voidaan ilmaista FWER=$P(a \geq 1)$. [3] On huomattava siis, että näin määriteltynä FWER tarkoittaa todennäköisyyttä hylätä yksi tai useampi tosi nollahypoteesi kaikkien hypoteesien populaatiossa.

	Päätös H_0		
	Hyväksytään	Hylätään	
Tosi H_0	$N_0 - a$	a	N_0
Epätosi H_0	$N_1 - b$	b	N_1
	$N - R$	R	N

Taulukko 1: Kunkin testin tulos voidaan sijoittaa yhteen ylläolevan taulukon soluista. Hylättyjen nollahypoteesien lukumäärästä R, virheellisten päätelmien lukumäärä on a. Oikeiden päätelmien lukumäärä on b.

Benjamini & Hochberg [3] osoittivat, että FDR:ää eli tosien nollahypoteesien odotettua lukumäärää kaikkien merkitseväksi valittujen nollahypotee-

sien joukossa, voidaan rajoittaa suureella q. Merkitsevyystason α avulla taas pyritään rajoittamaan tosien nollahypoteesien hylkäämistä. Olkoon $\alpha = q = 0.05$, tällöin

- α: 5% tosista nollahypoteeseista hylätään
- q-arvo: 5% kaikista hylätyistä nollahypoteeseista on tosia.

4.1 FDR-määritelmä

Virheellisesti hylättyjen nollahypoteesien osuutta kaikkien hylättyjen nollahypoteesien joukossa merkitään suureella $Q = a/(a+b)$, missä taulukon (1) merkintöjä käyttäen

$$a = \text{Hylättyjen tosien nollahypoteesien lukumäärä}$$

ja

$$b = \text{Hylättyjen epätosien nollahypoteesien lukumäärä}$$

Kun yhtään hylättyä nollahypoteesia ei ole, voidaan määritellä $Q = 0$, jos $a+b = 0$. Koska lukumäärää a ei tunneta, on myös hylkäysvirheiden (virheellisten löydösten) osuus Q tuntematon satunnaismuuttuja. Hylkäysvirheaste määritellään seuraavasti:

$$FDR = E[Q] = E[a/(a+b)] = E[a/R]. \tag{23}$$

FDR kertoo virheellisten löydösten odotusarvoisen osuuden kaikista hylätyistä nollahypoteeseista. Monivertailutilanteissa tunnusluvun FDR kontrollointi on noussut tärkeäksi työkaluksi tulosten raportoimisessa, sillä suurten aineistojen tapauksessa virheellisten löydösten osuuden rajoittaminen on informatiivisempaa kuin tyypin I virheen esiintyvyyden rajoittaminen.

FDR-menetelmällä on kaksi tärkeää ominaisuutta:

1. Kun yhtään hylkäystä ei tehdä, $b = 0$ ja $a = R$. Jos $a = 0$, niin tällöin $Q = 0$ ja kun $a > 0$ niin $Q = 1$, josta seuraa, että

   $$FWER = P(a \geq 1) = E[a/R] = E[Q] = FDR. \tag{24}$$

   Jos kaikki nollahypoteesit ovat tosia, FDR on siis ekvivalentti FWER:n kanssa.

2. Jos hylättäviä nollahypoteeseja löytyy, eli $N_0 < N$, FDR on pienempi tai yhtäsuuri kuin FWER. Jos virheellisiä päätöksiä on tehty, eli $a > 0$, mistä seuraa $a/R \leq 1$, jolloin

$$FWER = P(a \geq 1) \geq FDR \tag{25}$$

Jokainen prosessi, joka kontrolloi FWER:ää, kontrolloi siis myös FDR:ää.

4.1.1 FDR-kontrollointi

Olkoon N testattavien hypoteesien lukumäärä ja $p_1, p_2, ..., p_N$ testeistä saadut p-arvot suuruusjärjestyksessä. Kun p-arvot ovat toisistaan riippumattomia nollahypoteesin vallitessa, Benjamini & Hochberg osoittivat [3], että virheellisten löydösten osuutta voidaan kontrolloida suureen q avulla valitsemalla merkitsevät p-arvot seuraavasti:

$$p_i \leq \frac{i}{N} q, \tag{26}$$

missä q on ennaltamääritetty suure, joka voi saada arvoja väliltä $(0, 1)$. Olkoon i_{max} suurin luku joukosta $i \in \{1, ..., N\}$, joka toteuttaa yhtälön (26). Benjaminin & Hochbergin mukaan kaikki p-arvot p_i, joille $i \leq i_{max}$ valitaan merkitseväksi ja vastaavasti nollahypoteesit H_{0i}, joille $i \leq i_{max}$, hylätään. P-arvojen ollessa riippumattomia, Benjaminin & Hochbergin FDR-valintamenetelmä pyrkii rajoittamaan virheellisten löydösten osuutta suureen q avulla. Pätee, että

$$FDR = E[Q] \leq \frac{N_0}{N} q \leq q. \tag{27}$$

Kuten aiemmin todettu, monivertailun tapauksessa tosien nollahypoteesien lukumäärä N_0 on suuri, jolloin $N_0/N \approx 1$. Tyypillinen valinta on $q = 0.1$ [1].

4.1.2 Simulointeja

Benjaminin ja Hochbergin esittämää FDR-valintamenetelmän havainnoillistamista varten generoidaan kaksi toisistaan riippumatonta otosta, joissa $n = 10$, jakaumasta $\mathcal{N}(3, 1)$. Tarkastellaan simuloinnin tuloksia kolmella erilaisella lähestymistavalla:

1) Suoritetaan otoksien välille $N = 10000$ t-testiä

2) Muutetaan toista otosta niin, että enää 90% nollahypoteeseista on tosia. Havainnoista 10% noudattaa tällöin jakaumaa $\mathcal{N}(4, 1)$.

3) Simuloidaan $N = 10000$ samanaikaista t-testiä 1000 kertaa.

Olkoon nollahypoteesina nyt, että näiden kahden ryhmän odotusarvot ovat samat, jolloin sopiva testi tähän tapaukseen on kaksisuuntainen t-testi. Asetetaan merkitsevyystasoksi $\alpha = 0.05$ ja tehdään 10000 samanaikaista t-testiä ryhmille. Testien tuloksena hylätään yhteensä 487 nollahypoteesia. Koska molemmat otokset tulevat samasta jakaumasta $\mathcal{N}(3, 1)$, on virheellisten löydösten osuus 100%, sillä $Q = a/(a + b) = 487/(487 + 0) = 1$, mikä ei ole missään määrin hyväksyttävä lopputulos.

Muutetaan seuraavaksi toista otosta niin, että 10% havainnoista on epätosia nollahypoteseja, ja suoritetaan jälleen $N = 10000$ t-testiä. Tulokset ovat esitettynä taulukossa 2.

Nollahypoteesi	H_0 hyväksytään	H_0 hylätään	
Tosi	$N_0 - a = 8565$	$a = 435$	$N_0 = 9000$
Epätosi	$N_1 - b = 403$	$b = 597$	$N_1 = 1000$
	$N - R = 8968$	$R = 1032$	$N = 10000$

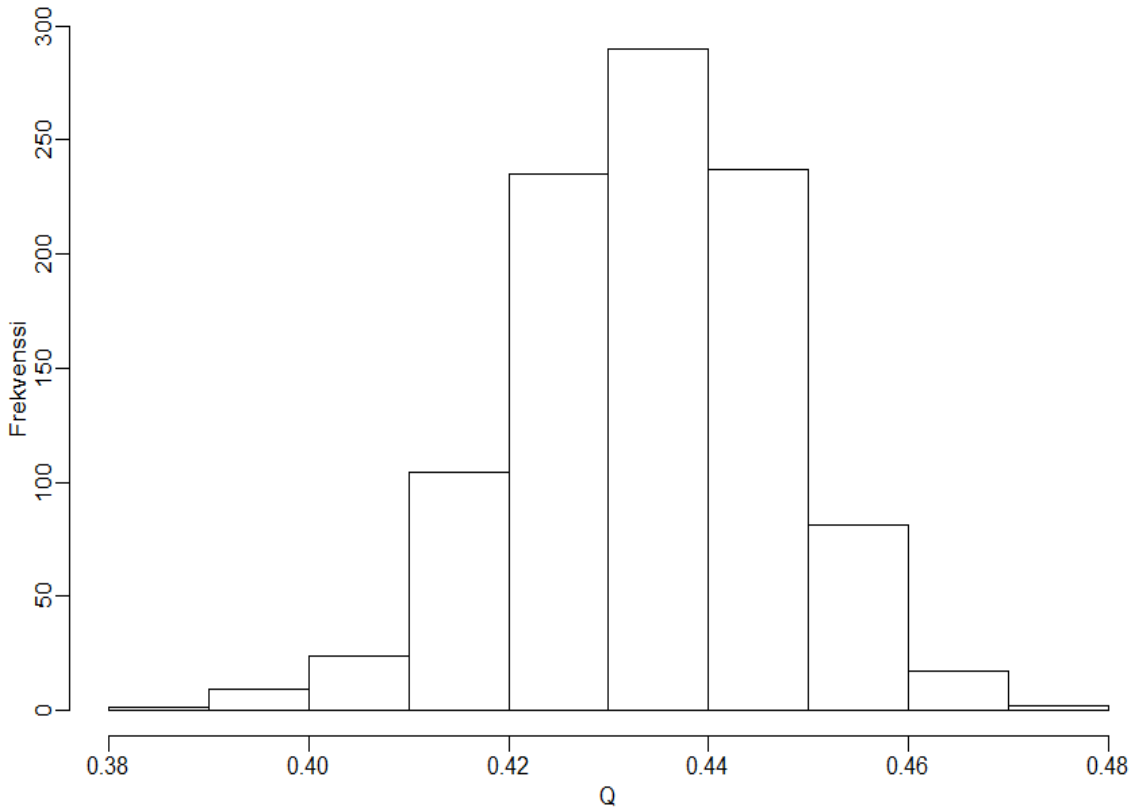
Taulukko 2: Nollahypoteesien hylkäämisien tulokset.

Taulukosta 2 voidaan laskea virheellisten löydösten osuus $Q = 435/1032 = 0.422$, mikä on suhteellisen korkea, sillä 90% nollahypoteeseista oli tosia. Etenkin geeniaineistoille on tyypillistä, että tosien nollahypoteesien osuus on todella suuri, minkä vuoksi on FDR:än kontrolloiminen on tärkeää.

Simuloidaan seuraavaksi 10000 t-testiä 1000 kertaa ja lasketaan kunkin simulaatiokierroksen tuottama Q. Simulaation tulokset on esitetty kuvassa 4.

Simulaatioiden perusteella voidaan laskea $FDR = \bar{Q} = 0.434$. FDR:än arvo tässä tapauksessa on suhteellisen korkea, joten korjataan sitä aiemmin esitellyllä Benjaminin & Hochbergin menetelmällä [3].

Toistetaan uudelleen edellinen prosessi, joissa toista otosta on muunnettu niin, että 90% nollahypoteeseistä on tosia:



Kuva 4: Histogrammi 1000:sta $N = 10000$ t-testin simulaatiosta.

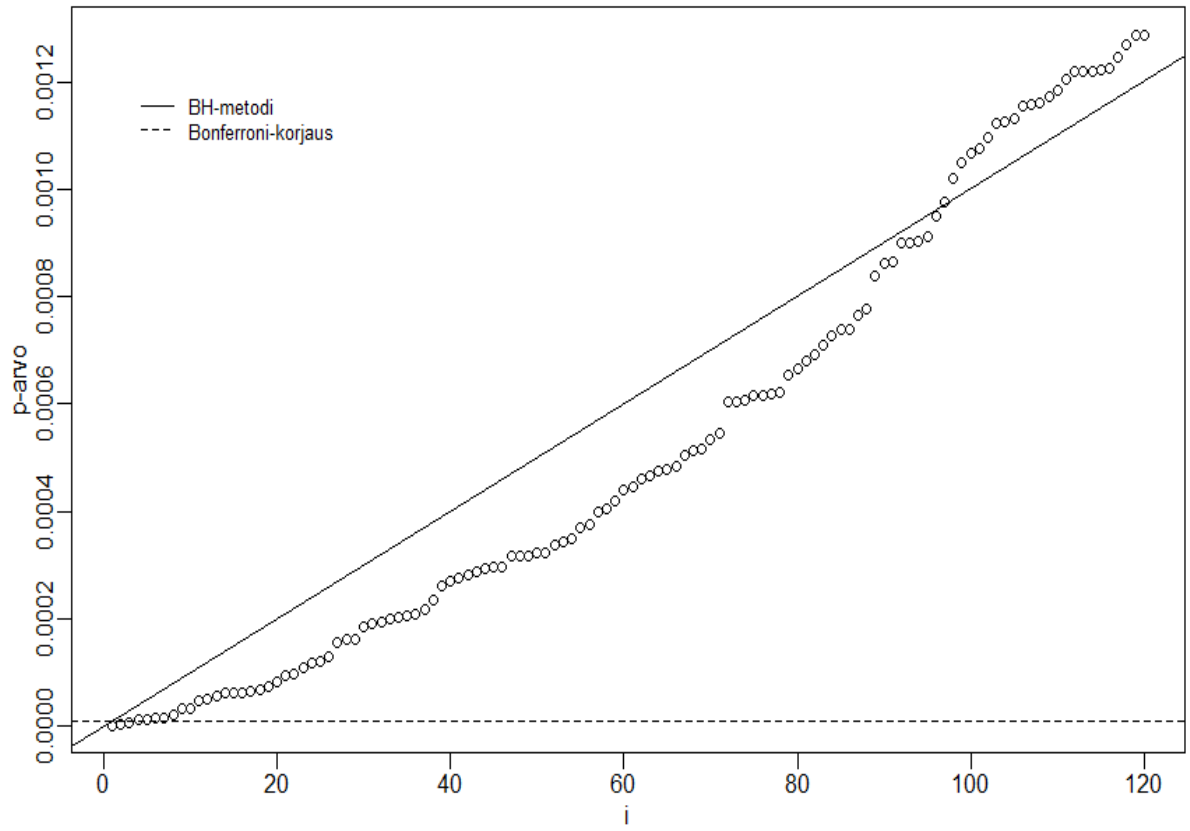
- Valitaan merkitsevyystasoksi $\alpha = 0.05$ ja kynnysarvoksi $q = 0.1$
- Tehdään $N = 10000$ samanaikaista kahden otoksen välistä vertailua.
- Järjestetään p-arvot suuruusjärjestykseen.
- Valitaan merkitseväksi ne p-arvot p_i, joille $i \leq i_{max}$, missä $p_{i_{max}}$ on suurin p-arvo, jolle pätee $p_i \leq \frac{i}{N}q$
- Toistetaan edelliset askeleet 1000 kertaa, jolloin kullekin simulaatiokierrokselle voidaan laskea virheellisten löydösten osuus $Q = a/(a + b)$
- Lasketaan FDR, joka on keskiarvo saaduista arvoista Q.

Kuvassa 5 on esitettynä yhden simulaatiokierroksen tuottamat 120 pienintä p-arvoa. Samaan kuvaan on myös vedetty suora, jonka kulmakerroin on qi/N. Kaikki suoran alapuolelle jääneet p-arvot valitaan merkitseväksi BH:n valintamenetelmän perusteella.

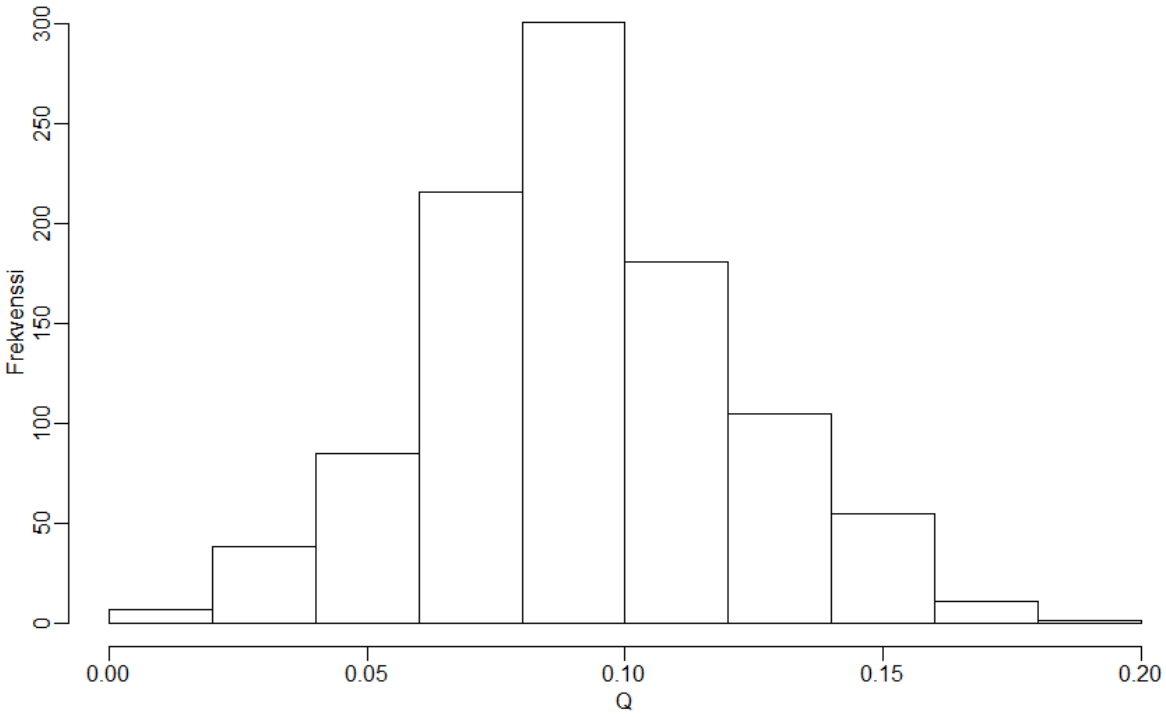
Bonferroni-korjauksen kanssa merkitseväksi valitaan kaikki ne p-arvot, jotka ovat pienempiä kuin $\alpha/N = 0.05/10000 = 0.000005$, joita löytyi yhteensä 2. Kuvassa 5 nämä ovat katkoviivan alle jäävät p-arvot. FDR-menetelmä valitsee enemmän nollahypoteeseja merkitseväksi, pitäen samalla virheellisten löydösten osuuden halutulla tasolla $q = 0.1$.

Simuloidaan vielä uudelleen 1000 kertaa $N = 10000$ t-testin tulosta, jotta voidaan varmistua, että FDR pysyy todellakin halutulla tasolla q. Simulaation tuottamat Q-arvot on esitettynä kuvassa 6. FDR:ksi saadaan $FDR = \bar{Q} = 0.0917$, joka todellakin on pienempi, kuin asetettu kynnysarvo $q = 0.1$, joten epäyhtälö (27) toteutuu.

FDR-menetelmä on nostanut suosiotaan, kun halutaan hallita tyypin I virheen esiintyvyyttä, sille se on hiukan sallivampi kuin edellä esitetty FWER-menetelmä [1].



Kuva 5: BH:n FDR-valintamentelmän perusteella, merkitseväksi valitaan 95 p-arvoa (yhtenäisen viivan alapuolelle jäävät arvot), kun taas Bonferroni-korjauksen perusteella hylätään vain 2 nollahypoteesia (katkoviivan alle jäävät arvot)



Kuva 6: Simulaatio suoritettuna 1000 kertaa $N = 10000$ t-testille.

5 Bayesiläinen monivertailu

Tässä luvussa määritellään tunnusluku FDR bayesiläisen päättelyn kautta, kuten Efron [8] on esittänyt alun perin. Efron myös osoitti, että FDR perustuu nollahypoteesien (eli tämän tutkielman esimerkkiaineiston nollageenien) posterioritodennäköisyyksille. Määritellään syöpäaineiston tapauksessa ensin prioritodennäköisyydet π_0 sekä $\pi_1 = 1 - \pi_0$. Jokainen geeni i on siis joko nollageeni tai ei-nollageeni vastaavilla prioritodennäköisyyksillä. Lasketaan aineistolle z-arvot lausekkeen (16) mukaan. Oletetaan jälleen, että z-arvot noudattavat normaalijakaumaa, $z_i \backsim \mathcal{N}(\mu_i, 1)$. Geeni i on nollageeni, jos $\mu_i = 0$.

Muunnoksen kautta saavutetuilla z-arvoilla on tiheysfunktiot $f_0(z)$ ja $f_1(z)$. Koska usein monivertailussa tosien nollahypoteesien määrä on suuri, todennäköisyys π_0 on lähellä ykköstä. Nollahypoteesin mukaan $H_{0i} : z_i \backsim \mathcal{N}(0, 1)$, jolloin tiheysfunktio $f_0(z)$ on standardoidun normaalijakauman tiheysfunktio:

$$f_0(z) = \phi(z) = \exp(-\frac{1}{2}z^2)/\sqrt{2\pi}. \tag{28}$$

Vastaavasti $f_1(z)$ on edelleen estimoitavissa ja se voi saada arvonsa jostakin toisesta tiheysfunktiosta [8]. Olkoon $\mathcal{Z}$ mikä tahansa osajoukko reaaliluvuista $\mathbb{R}$. Yksisuuntaisten testien tapauksessa $\mathcal{Z} : Z \leq z$ tai $\mathcal{Z} : Z \geq z$, riippuen testin suunnasta. Kaksisuuntaisen testin tapauksessa $\mathcal{Z} : Z \geq |z|$. Jakaumafunktiot voidaan määritellä

$$F_0(\mathcal{Z}) = \int_{\mathcal{Z}} f_0(z)dz \text{ ja } F_1(\mathcal{Z}) = \int_{\mathcal{Z}} f_1(z)dz. \tag{29}$$

Olkoon sekoitusjakauman tiheysfunktio muotoa

$$f(z) = \pi_0 f_0(z) + \pi_1 f_1(z), \tag{30}$$

jolloin sekoitusjakauman kertymäfunktio voidaan määritellä:

$$F(\mathcal{Z}) = \pi_0 F_0(\mathcal{Z}) + \pi_1 F_1(\mathcal{Z}). \tag{31}$$

Bayesin kaavan mukaan nyt voidaan määritellä nollageeneille posterioritodennäköisyys

$$\Phi(\mathcal{Z}) = P(\text{geeni } i \text{ on nollageeni} | z \in \mathcal{Z}) \tag{32}$$

$$= \pi_0 F_0(\mathcal{Z})/F(\mathcal{Z}). \tag{33}$$

Todennäköisyyttä $\Phi(\mathcal{Z})$ voidaan kutsua bayesiläiseksi hylkäysvirheasteeksi, jota tästä eteenpäin merkitään tunnusluvulla $\overset{Bayes}{Fdr}$. Jos havainto z raportoidaan nollageeniksi kun $z \in \mathcal{Z}$, on $\Phi(\mathcal{Z})$ todennäköisyys, että kyseessä on virheellinen löydös.

5.1 Empiirinen määritelmä

Kuten aikaisemmin on mainittu, tosien nollahypoteesien, tässä tapauksessa nollageenien osuus, on lähellä yhtä. Toistaiseksi määritellään siis $\pi_0 = 1$. Oletus myös nollahypoteesin mukaisesta jakaumasta säilyy, eli $H_{0i} : z_i \backsim \mathcal{N}(0, 1)$. Efron [8] määritteli empiirisen estimaatin jakaumafunktiolle $F(\mathcal{Z})$:

$$\bar{F}(\mathcal{Z}) = \#\{z_i \in \mathcal{Z}\}/N, \tag{34}$$

eli $\bar{F}(\mathcal{Z})$ kertoo niiden z-arvojen, jotka kuuluvat tarkasteltavaan osajoukkoon $\mathcal{Z}$, osuuden kaikista z-arvoista. Yhtälön (31) perusteella voidaan kirjoittaa empiirinen estimaatti FDR:lle

$$\overline{Fdr}(\mathcal{Z}) = \pi_0 F_0(\mathcal{Z})/\bar{F}(\mathcal{Z}). \tag{35}$$

Kun N on suuri, $\overline{F}(\mathcal{Z}) \approx F(\mathcal{Z})$ jolloin $\overline{Fdr}$ on hyvä estimaatti bayesiläiselle Fdr:lle [8].

Oletetaan, että Benjaminin ja Hochbergin FDR-valintamenetelmän p-arvot p_i vastaavat reaaliarvoisia z-arvoja z_i siten, että $p_i = F_0(z_i)$. Tällöin tarkasteltavana on yksisuuntainen t-testi, jolloin osajoukko $\mathcal{Z}$, eli kriittinen alue, määritellään $\mathcal{Z} : Z \leq z$. Kuten Benjaminin ja Hochbergin valintamenetelmässä, järjestetään z-arvot suuruusjärjestykseen

$$z_{(1)} \leq z_{(2)} \leq ... \leq z_{(N)}. \tag{36}$$

Nyt yhtälö (34) voidaan kirjoittaa

$$\bar{F}(z) = \#\{z_i \leq z\}/N, \tag{37}$$

jolloin i:nnelle z-arvolle saadaan

$$\bar{F}(z_{(i)}) = i/N. \tag{38}$$

Huomioitavaa on, että lauseke (37) voitaisiin ilmaista myös yksisuuntaisen testin tapauksessa

$$\bar{F}(z) = \#\{z_i \geq z\}/N, \tag{39}$$

kun $\mathcal{Z} : Z \geq z$. Tässä tutkielmassa tarkastellaan kuitenkin tästä eteenpäin yksisuuntaista testiä, kun $\mathcal{Z} : Z \leq z$.

Benjaminin ja Hochbergin määrittelemä hylkäysraja (26) voidaan ilmaista bayesiläisittäin nyt muodossa:

$$p_i \leq \frac{i}{N} q = \bar{F}(z_{(i)}) q, \tag{40}$$

eli

$$F_0(z_{(i)}) \leq \bar{F}(z_{(i)}) q, \tag{41}$$

josta seuraa edelleen

$$F_0(z_{(i)}) / \bar{F}(z_{(i)}) \leq q, \tag{42}$$

tai yleisemmin, jos oletetaan, että $\pi_0 \neq 1$, niin

$$\pi_0 F_0(z_{(i)}) / \bar{F}(z_{(i)}) \leq \pi_0 q. \tag{43}$$

Oletetaan, että $\pi_0 = 1$. Olkoon i_{max} suurin indeksi, jolle pätee

$$\overline{Fdr}(z_{(i)}) \leq q. \tag{44}$$

Edellisten kaavojen perusteella tämä tarkoittaa, että i_{max} on sama kuin Benjaminin ja Hochbergin valintamenetelmän mukainen i_{max}.

Efron [8] todisti empiisellä bayesiläisellä estimaattorilla olevan lähes välitön yhteys Benjaminin & Hochbergin hylkäysmenetelmään, sillä bayesiläinen hylkäyssääntö julistaa ne geenit ei-nollaksi, joille $z_i \leq z$ ja kontrolloi FDR:ää tasolla q:

$$\overset{Bayes}{Fdr}(z) < q \tag{45}$$

Bayesiläinen posterioritodennäköisyys nollageeneille on siis Benjaminin & Hochbergin määrittelemä FDR-valintamenetelmä, joten voidaan todeta, että FDR:än kontrollointi perustuu tosiasiassa todennäköisyyksien laskemiseen.

Tarkastellaan seuraavaksi bayesiläisen FDR:n laskemista esimerkkiaineiston avulla. Syöpäaineiston $N = 6033$ ja valitaan kynnysarvoksi $q = 0.1$. Oletetaan jälleen, että $\pi_0 = 1$. Maksimi i_{max}, joka toteuttaa yhtälön

$$F_0(z_{(i)}) / (i/N) < q, \tag{46}$$

on $z = -3.282$, jonka indeksi $i = 32$. Nollahypoteeseja hylätään siis 32 kappaletta, ja $\overset{Bayes}{Fdr}$ on enintään $q = 0.1$.

Tarkastellaan vielä bayesiläistä FDR:ää kuvan 7 mukaisesti. Merkitään tarkasteltavaa kynnysarvoa z_0, jolloin $N(z_0) = \#\{z_i \leq z_0\}$. Nollahypoteesi hylätään tai hyväksytään nyt seuraavin ehdoin:

$$\begin{cases} H_0, & z_i > z_0. \\ H_1, & z_i \leq z_0. \end{cases} \tag{47}$$

Taulukon 1 perusteella nollageeneille $N_0(z_0) = a$ kun z-arvot alittavat kynnysarvon z_0 ja vastaavasti ei-nollageeneille $N_1(z_0) = b$, jolloin

$$N(z_0) = N_0(z_0) + N_1(z_0) = a + b = R. \tag{48}$$

Määritellään virheellisten löydösten osuus (vrt. Q lausekkeessa (23))

$$Fdp = \frac{N_0(z_0)}{N(z_0)}. \tag{49}$$

Kuvassa 7 havainnoillistetaan havaintoarvojen jakautumista nollageeneihin ja ei-nollageeneihin prioritodennäköisyyksillä π_0 ja π_1. Ainoastaan sekoitusjakauman tiheysfunktion $f(z) = \pi_0 f_0(z) + \pi_1 f_1(z)$ tuottama arvo z_i havaitaan, ja nollahypoteesi H_{0i} hylätään, jos z_i alittaa (oikeanpuoleisten arvojen tarkastelussa ylittää) valitun kynnysarvon z_0.

Kuten Benjaminin & Hochbergin FDR-valintamenetelmässä, hylättyjen ja tosien nollahypoteesien lukumäärän a tarkkaa arvoa ei tunneta, mutta sen odotusarvo on:

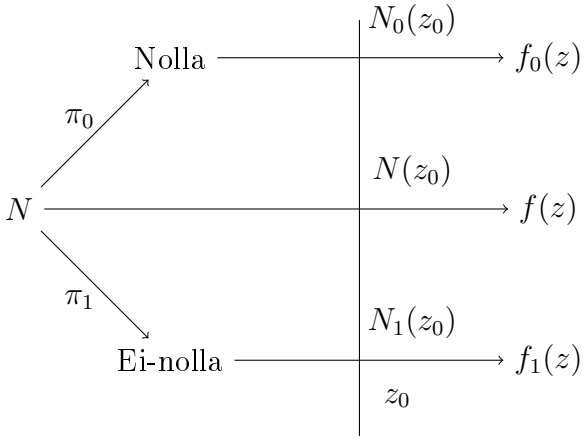
$$E[a] = E[N_0(z_0)] = N\pi_0 F_0(z_0). \tag{50}$$

Määritelmä (35), voidaan ilmaista nyt

$$\overline{Fdr}(z_0) = N\pi_0 F_0(z_0)/N(z_0), \tag{51}$$

liittämällä yhtälöön (51) empiirinen estimaatti jakaumafunktiolle, kts. yhtälö (37), saadaan suoraan kirjoitettua baysiläisen hylkäysvirheasteen määritelmä:

$$\overline{Fdr}(z_0) = N\pi_0 F_0(z_0)/N(z_0) = \frac{\pi_0 F_0(z_0)}{\overline{F}(z_0)} = \overset{Bayes}{\widehat{Fdr}}(z_0). \tag{52}$$



Kuva 7: Diagrammi sekoitusjakauman tiheysfunktion muodostamisesta. Kuva on muunnelma viitteen [1] kuvasta 15.4.

Syöpäaineistossa $N = 6033$, joista $N(z_0) = 41$ z-arvoa on alittanut ennaltamääritellyn kynnysarvon $z_0 = -3$. Nyt $F_0(z_0) = \Phi(-3)$ sekä $\pi_0 = 1$ nollahypoteesin vallitessa, jolloin

$$E[N_0(z_0)] = 6033 * 1 * (\Phi(-3)) = 8.14. \tag{53}$$

Hylkäysvirheasteen estimaatiksi saadaan

$$\overline{Fdr}(z_0) = 8.14/41 = 0.199, \tag{54}$$

eli esimerkin tapauksessa voidaan sanoa, että noin 20% näistä 41 merkitseviksi valituista nollageenistä oli virheellisiä löydöksiä.

5.2 Lokaali FDR

Aiemmissa kappaleissa esitetty FDR-menetelmä kontrolloi globaalia tunnuslukua FDR, mikä on yleisimmin käytetty lähestymistapa. Globaalin tunnusluvun FDR tarkastelu perustuu testisuureen häntäjakaumaan. Tietyissa tutkimuksissa voidaan kuitenkin haluta tarkastella yksittäisiä poikkeavia havaintoja. On mahdollista, että ennen tutkimuksen suorittamista tiedetään, että jokin tietty geeni käyttäytyy useimmiten eri tavalla kuin joukko muita geenejä. Tai vaihtoehtoisesti esimerkiksi kahden otoksen tutkimuksessa, jossa tutkijat haluavat tietää, mitkä ovat ne tietyt geenit joukosta $N = n_1 + n_2$, jotka poikkeavat toisistaan. Globaali FDR löytää poikkeavat geenit, mutta entä jos näiden tiettyjen geenin käyttäytymistä halutaan tutkia tarkemmin?

Olkoon π_0 ja π_1 sekä $f_0(z)$ ja $f_1(z)$ nollageenejä ja ei-nollageenejä vastaavat prioritodenäköisyysdet ja tiheysfunktiot. Merkittävin ero globaalin ja lokaalin hylkäysivhreasteen tarkastelussa on se, että lokaali FDR perustuu tiheysfunktioiden tarkasteluun, kun taas globaali FDR perustuu häntätodennäköisyyksien tarkasteluun.

B. Efron [8] määritteli lokaalin FDR:än tiheysfunktioiden avulla

$$fdr(z_0) = P(\text{geeni } i \text{ on nollageeni}|z_i = z_0) = \pi_0 f_0(z)/f(z), \tag{55}$$

jossa $f(z)$ on aikaisemmin määritelty sekoitusjakauman tiheysfunktio $f(z) = \pi_0 f_0(z) + \pi_1 f_1(z)$. Oletetaan jälleen, että $\pi_0 = 1$. Samoin nollahypoteesin mukainen tiheysfunktio $f_0(z)$ tunnetaan. Olkoon aiemman mukaan $f_0(z) = \phi(z) = \exp(-\frac{1}{2}z^2)/\sqrt{2\pi}$. Jos tarkasteltavana olisi vain yksi geeni i, nollahypoteesi hylättäisiin tasolla $\alpha = 0.05$, jos $|z_i| > 1.96$. Näin ollen ainoaksi estimoitavaksi suureksi jää $f(z)$.

Geeniä i vastaava estimaatti lokaalille FDR:lle on

$$\widehat{fdr}(z_i) = \frac{\hat{\pi}_0 f_0(z_i)}{\hat{f(z_i)}} \tag{56}$$

missä voidaan asettaa $\hat{\pi}_0 = 1$. Efronin mukaan yleinen tapa raportoida kiinnostavat havainnot, tässä tapauksessa geenit, on asettaa $\widehat{fdr}(z_i) \leq 0.20$.

5.2.1 Poisson-regression estimaatit $f(z)$:lle

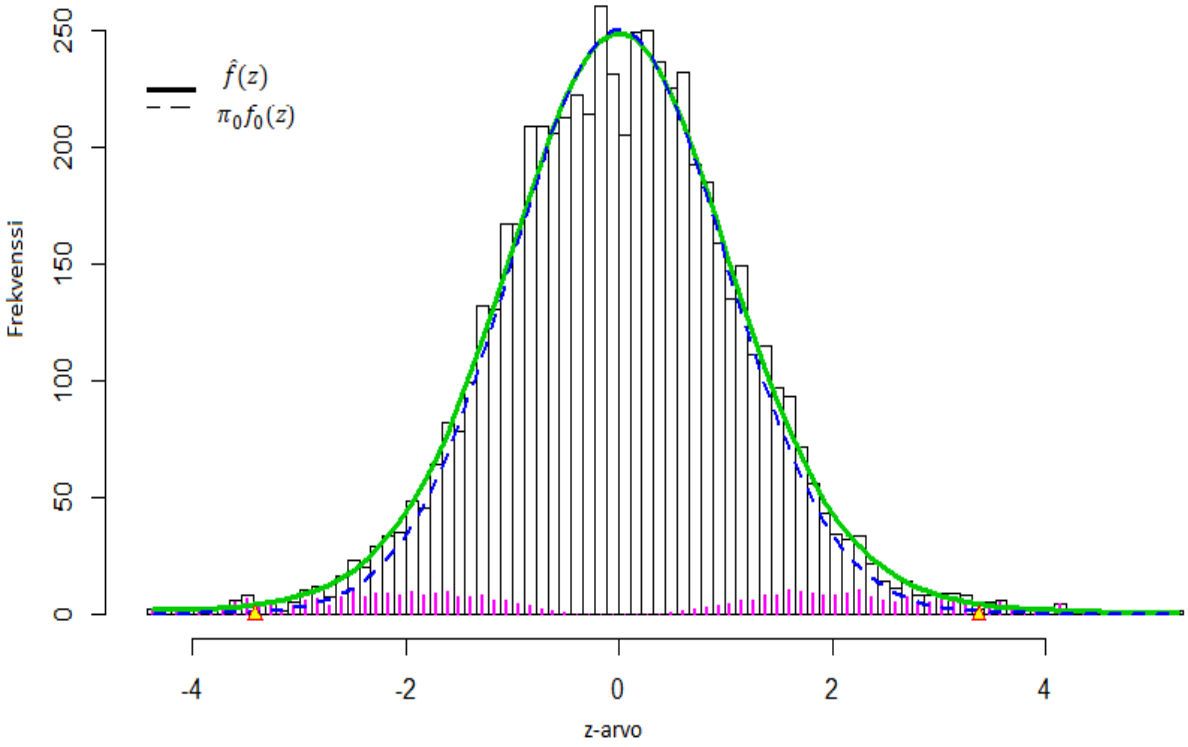
Sekoitusjakauman tiheysfunktio $f(z)$ voidaan estimoida käyttäen testisuureen $\mathbf{z}$ arvoja $\mathbf{z} = \{z_1, ..., z_n\}$. Efronin mukaan [9] tarpeeksi hyvä estimaatti $\hat{f(z)}$ saadaan diskretoiduille z-arvoille Poisson-regression suurimman uskottavuuden estimoinnin periaatteita noudattaen. Sekoitusjakauma tiheysfunktiota $f(z)$ estimoidaan niiden empiiriseen jakaumaan sovitetun Poisson-mallin avulla.

Olkoon z-arvot jaettu osajoukkoihin k, joissa jokaisen osajoukon pituus, eli z-akselin vaihteluväli on d. Huomioitavaa on, että välin pituuden d täytyy olla yhtä suuri jokaiselle osajoukolle, jolloin välin pituus d vaihtelee sen perusteella, kuinka moneen osajoukkoon k z-arvot on jaettu. Merkitään

$$y_k = \#\{z_i \text{ osajoukossa } k\},\ k = \{1, ..., K\} \tag{57}$$

ja

$$x_k = \text{osajoukon } k \text{ keskipiste} \tag{58}$$



Kuva 8: Syöpäaineiston havainnot jaettuna $K = 89$ osajoukkoon, jolloin välin pituus $d = 0.1$. Pystysuorat viivat palkkien sisällä kuvaavat ei-nollageenien estimoitua lukumäärää. Kolmiolla merkittynä raja $\widehat{fdr}(z_i) = 0.20$

Syöpäaineiston tapauksessa aineisto on jaettu $K = 89$ osajoukkoon, mistä seuraa, että jokaisen osajoukon pituus $d = 0.1$. Tällöin $x_1 = -4.45, x_2 = -4.35, ..., x_{89} = 4.45$. Olkoon z-arvojen odotettu kokonaislukumäärä osajoukossa k

$$E(y_k) = \nu_k = N * d * f(x_k). \tag{59}$$

Efronin mukaan [9] lukumäärien y_k voidaan olettaa olevan riippumattomasti Poisson-jakautuneita,

$$y_k \backsim Po(\nu_k), k = 1, ..., K. \tag{60}$$

Mallintamalla parametria $log(\nu_k)$ keskipisteen x_k p:nnen asteen polynomisella funktiolla, voidaan muodostaa Poissonin log-lineaarinen malli, jolloin

$$log(\nu_k) = \sum_{j=0}^{J} \beta_j x_k^j. \tag{61}$$

Odotusarvon sovitettu käyrä on siis verrannollinen sekoitusjakauman tiheysfunktioon $f(z)$.

5.2.2 Lokaalin FDR:n tulkinta

Ohjelmasta R löytyy paketti locfdr lokaalin fdr:än analysoimiseen. Funktio locfdr etsii sopivimman estimaatin funktiolle $\hat{f}(z)$ käyttäen edellä esitettyä menetelmää. Kuvassa 8 on esitettynä estimoitu $\hat{f}(z)$. Pystysuorat viivat histogrammin osajoukkojen sisällä kuvaavat estimoituja ei-nollageenien lukumäärää.

Merkitään nollageenien odotettua lukumäärää osajoukossa k:

$$e_0(k) = d * N * f_0(x_k). \tag{62}$$

Efronin [8] mukaan lokaali fdr-estimaatti voidaan nyt kirjoittaa muodossa

$$\widehat{fdr} = e_0(k)/y_k. \tag{63}$$

Eturauhaussyöpädatan $N = 6033$ geenistä, yhteensä 51 z-arvoa toteuttaa Efronin määrittelemän ehdon $\hat{fdr}(z_i) \leq 0.20$. Kuvassa 8 nämä ovat siis ne arvot, jotka ovat pienempiä kuin $z = -3.415$ (merkittynä vasemmanpuoleisella kolmiolla kuvassa 8) tai suurempia kuin $z = 3.371$ (merkittynä oikeanpuoleisella kolmiolla kuvassa 8). Z-arvoja, jotka toteuttavat ehdon $z_i \leq -3.415$ on yhteensä 26 kpl. Vastaavasti z-arvoja, jotka toteuttavat ehdon $z_i \geq 3.371$ on

25 kpl. Tarkastellaan lähemmin kynnysarvoa $z = 3.371$. Tästä arvosta seuraava osajoukko on $[3.4, 3.5)$, joka on järjestysluvultaan osajoukko $k = 80$ ja välin keskipiste $x_{80} = 3.45$. Tämä osajoukko sisältää yhteensä neljä z-arvoa, $y_{80} = 4$. Nyt yhtälö (62) voidaan ilmaista:

$$e_0(80) = 0.1 * 6033 * \phi(3.45), \tag{64}$$

jolloin lokaaliksi esimaattoriksi saadaan

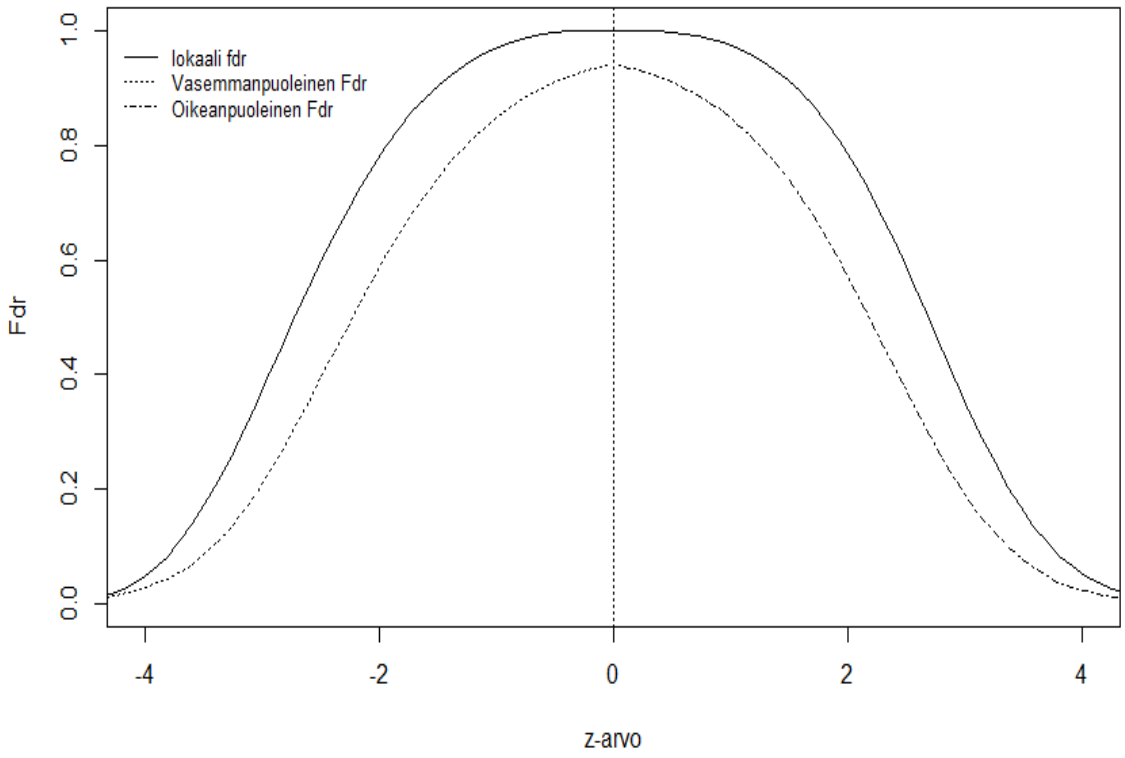
$$\widehat{fdr}_{80} = e_0(80)/y_{80} = 0.626/4 = 0.157. \tag{65}$$

Syöpäaineiston tapauksessa lokaali fdr on konservatiivisempi, sillä hylättyjä nollahypoteeseja on vähemmän. Oikeanpuoleisen yksisuuntaisen testin tapauksessa Benjaminin ja Hochbergin bayesiläisellä menetelmällä hylättyjä z-arvoja on 28 ja vasemmanpuoleisen yksisuuntaisen testin tapauksessa hylättyjä z-arvoja on 32. Kuvassa 9 on esitettynä syöpäaineiston lokaali ja globaali FDR.

Efron todisti [9] lokaalilla ja globaalilla Fdr:llä olevan välitön yhteys toisiinsa:

$$Fdr(z) = E_f[fdr(z)|Z \leq z], \tag{66}$$

missä F_f on odotusarvo $f(z)$ suhteen. Huomioitavaa on, että edellä esitetyssä lausekkeessa epäyhtälö $Z \leq z$ voitaisiin myös korvata epäyhtälöllä $Z \geq z$ tai kaksisuuntaisen testin tapauksessa epäyhtälöllä $Z \geq |z|$.



Kuva 9: Syöpäaineiston lokaali FDR ja yksisuuntaisten testien $\overset{Bayes}{Fdr}$

6 Muunnelmat BH:n FDR-valintamenetelmästä

Benjaminin & Hochbergin valintamenetelmälle on kehittetty lukuisia hieman konservatiivisempia muunnelmia. Tässä tutkielmassa näistä esitetään Benjaminin & Hochbergin kehittämä metodi [4] sekä Benjaminin & Yekutielin & Kriegerin kehitämä metodi [7]. Näitä menetelmiä kutsutaan adaptiivisiksi metodeiksi, sillä niissä merkitsevien p-arvojen valinta suoritetaan kaksi kertaa, tai jopa useammin. Perusideana adaptiivisissä prosesseissa on se, että ensimmäisen valinnan perusteella muodostetaan arvio tosien nollahypoteesien lukumäärästä N_0, jota käytetään arviona seuraavan kierroksen prosessissa, jolloin p-arvojen hylkäysraja tarkentuu.

Yleisesti adaptiiviset menetelmät suoritetaan seuraavalla tavalla:

1. Lasketaan uusi estimaatti $\hat{N_0}$ tosille nollahypoteeseille
2. Jos tosia nollahypoteeseja ei ole, lopetetaan. Muuten jatketaan p-arvojen arvioimista uudella tasolla $Nq/\hat{N_0}$

6.1 Benjaminin & Hochbergin adaptiivinen metodi

Benjamin & Hochberg [4] kehittivät adaptiivisen metodin, joka perustuu tosien nollahypoteesien lukumäärän N_0 estimoimiseen alimman tason käyrän (engl. *LSL-curve*) avulla.

Jos kaikki nollahypoteesit ovat tosia, eli $N_0 = N$, ja testisuureet ovat toisistaan riippumattomia, niin havaittuja p-arvoja voidaan pitää järjestettynä realisaationa jakauman $\mathcal{U}(0,1)$ otoksesta. Tällöin näiden p-arvojen odotusarvo voidaan ilmaista $E[p_i] = i/(N+1)$. P-arvojen kvantiilikuvio on tällöin suora, jonka kulmakerroin on vakio $S = 1/(N+1)$.

Kun joukossa on myös epätosia nollahypoteeseja eli $N_0 < N$, p-arvot, jotka vastaavat epätosia nollahypoteeseja, sijoittuvat kvantiilikuviossa vasemmalle puolelle. Tosia nollahypoteeseja vastaavat p-arvot sijoittuvat kvantiilikuviossa oikealle puolelle, sekä seuraavat approksimoidusti lineaarista käyrää $\beta = 1/(N_0+1)$. Sopivalla määrällä hyväksyttyjä p-arvoja, voidaan muodostaa kvantiilikuvion mukaisesti estimoitu käyrä kulmakertoimella $\hat{\beta}$, jolloin tosien nollahypoteesien määrän estimaatiksi saadaan $\hat{N_0} = 1/\hat{\beta}$.

Benjamin & Hochberg osoittivat, että sopiva metodi hyväksyttyjen p-arvojen valitsemiseen on niin kutsuttu LSL-metodi, jossa järjestettyjä p-arvoja p_i verrataan toisiinsa käyrän

$$S_i = (1 - p_i)/(n + 1 - i), \tag{67}$$

avulla. BH:n adaptiivinen prosessi suoritetaan seuraavasti:

1. Järjestetään p-arvot
2. Jos yksikään p_i ei toteuta $p_i \leq \frac{i}{N}q$, niin lopetetaan, ja kaikki nollahypoteesit pidetään voimassa
3. Lasketaan käyrät $S_i = (1 - p_i)/(n + 1 - i)$
4. Vertaillaan käyriä edelliseen aloittaen $i = 1$. Kun $S_i > S_{i-1}$, vertailu lopetetaan, ja estimaatiksi asetetaan

$$\hat{N}_0 = \min\{1/S_i + 1, N\}. \tag{68}$$

5. Toistetaan vaihe 2. aloittaen suurimmasta p-arvosta, ja etsitään ensimmäinen p-arvo $p_{i_{max}}$, joka toteuttaa yhtälön $p_i \leq \frac{i}{\hat{N}_0}q$.
6. Valitaan merkitseviksi p-arvoiksi kaikki ne, joille $i \leq i_{max}$

Benjamin & Hochberg osoittivat simulaatiotutkimusten avulla, että edellä esitetty adaptiivinen metodi kontrolloi FDR:ää tasolla q riippumattomuuden vallitessa, sekä on myös konservatiivisempi kuin alkuperäinen BH-metodi.

6.2 Benjaminin & Yekutielin & Kriegerin metodi

Benjamin & Yekutiel & Krieger [7] ehdottivat uutta adaptiivista metodia hylkäysrajan määrittämiseen. Menetelmä perustuu BH:n adaptiiviseen metodiin [4], jonka jälkeen suoritetaan vielä uusi askeltava tarkastelu hylkäysrajaan. Ensimmäisellä kierroksella valinta suoritetaan käyttäen kynnysarvoa $q' = q/(q + 1)$, jonka jälkeen toinen kierros suoritenaan käyttäen hyväksi ensimmmäisen kierroksen p-arvoja, jotka valittiin merkitseviksi

1. Suoritetaan BH-metodi käyttäen kynnysarvoa $q' = q/(q + 1)$
2. Olkoon r_1 hylättyjen nollahypoteesien lukumäärä. Jos $r_1 = 0$, lopetetaan ja kaikki nollahypoteesit hyväksytään. Jos taas $r_1 = N$, kaikki nollahypoteesit hylätään.
3. Lasketaan uusi estimaatti tosille nollahypoteeseille $\hat{N}_0 = (N - r_1)$

4. Toistetaan BH:n metodi; etsitään ensimmäinen p-arvo $p_{i_{max}}$, joka toteuttaa yhtälön
$$p_i \leq \frac{i}{\hat{N}_0} q'. \tag{69}$$

5. Valitaan merkitseviksi p-arvoiksi kaikki ne, joille $i \leq i_{max}$

Benjamin & Yekutiel& Krieger [7] todistivat, että BKY-metodi kasvattaa hylkäysrajan voimakkuutta, sekä kontrolloi FDR:rää tasolla q. Lisäksi simulaatiotutkimusten avulla todistettiin, että BKY-sallii tietynlaista positiivista korrelaatiota edelleen kontrolloidessaan FDR:rää.

7 Pohdinta

Jo tutkimusta suunniteltaessa on tärkeää huomioida monivertailuasetelma ja selvittää, minkälainen lähestymistapa on hyvä ottaa tulosten tarkasteluun. Ennen FDR-menetelmän kehittymistä suurin päätös on tehty tyypin I virheen kontrolloimisen kanssa. Jos lähtökohtana käytetään yksinkertaisesti tyypin I virheen esiintymistodennäköisyyden rajoittaminsta, turvaudutaan esimerkiksi Bonferroni-korjaukseen tai muuhun FWER:tä rajoittavaan metodiin. Tässä tutkielmassa on tarkasteltu menetelmiä, joiden mukaan hyvä vaihtoehto suurten ainestojen tapauksessa olisikin keskittyä hylkäysvirheasteen kontrolloimiseen, tyypin I virheen kontrolloimisen sijasta.

Esimerkiksi geeniainestojen tapauksessa, joita tässäkin tutkielmassa on esitelty, samanaikaisesti testattavina on kuitenkin tuhansia hypoteesipareja, jolloin Bonferroni-korjauksen määritelmän perusteella hylkäysraja pienenee sitä mukaa kuin samanaikaisesti suoritettavien testien lukumäärä kasvaa. Tuloksena on usein vain muutama merkitsevä löydös, mikä voi helposti johtaa vääriin loppupäätelmiin.

Muun muassa M.E Glickman ym. [13] suosittelevatkin FDR-menetelmän käyttämistä perinteisen p-arvojen korjausten sijaan. Kuitenkin toistaiseksi vielä p-arvojen korjaaminen on tutkijoiden joukossa suositumpi tapa, vaikka sen tuottamat tulokset ovatkin usein liian konservatiivisia. Tämä johtuu luultavasti siitä, että FDR-kontrollointi on monille tutkijoille vielä toistaiseksi tuntematonta [13]. Otoskoon kasvaminen ei myöskään vaikuta FDR-menetelmän toimivuuteen. P-arvojen korjausmenetelmissä yleisesti otoskoon kasvattaminen vaikuttaa konservatiivisesti hylkäämispäätöksiin, mutta FDR-menetelmän on todettu toimivan hyvin sekä isoilla että pienillä otoskoilla [13].

Vaikka FDR-menetelmä kehitettiin jo 1990-luvun alussa, vasta viime vuosina se on nostanut suosiotaan monivertailutestien yhteydessä, sillä vasta suurten aineistojen tapauksessa sen todellinen hyöty ja tehokkuus huomataan. Benjaminin & Hochbergin kehittämä FDR-valintamenetelmä on edelleen yksi suosituimmista keinosta hallita FDR:ää, vaikka siitä on kehitelty monenlaisia muunnelmia. BH-valintamenetelmän kehittämisen jälkeen, Benjamini ja Yekutieli todistivat nimittäin, että alkuperäinen menetelmä sallii myös tietynlaisen riippuvuuden testien välillä. [5]

Jo pelkän simulaation avulla todistettiin, että BH-valintamenetelmä todella löytää enemmän merkitseviä nollahypoteeseja kuin Bonferroni-korjaus mutta rajoittaa kuitenkin samalla hylkäysvirheastetta. Suurten aineistojen tapauk-

sessa onkin yleensä hyödyllisempää hallita hylkäysvirheastetta, kuin tyypin I virheen esiintymistodennäköisyyttä. Hylkäysvirheasteen avulla saadaan tutkijoiden kannalta mielekkäämpiä tuloksia, sillä hylkäysvirheasteen avulla pyritään rajoittamaan tosien nollahypoteesien liiallista hylkäämistä.

Viitteet

[1] B. Efron & T. Hastie *Computer Age Statistical Inference*, Cambridge University Press, 2016, U.S.A

[2] S. Dudoit & M. van der Laan *Multiple Hypothesis Testing. In: Multiple Testing Procedures with Applications to Genomics*, Springer Series in Statistics. Springer, 2008

[3] Y. Benjamini & Y. Hochberg *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society, Series B. vol. 57, no. 1, pp. 289–300, 1995

[4] Y. Benjamini & Y. Hochberg *On the Adaptive Control of the False Discovery Rate in Multiple Testing with Independent Statistics.* Journal of Educational and Behavioral Statistics, Vol. 25, No. 1, pp. 60-83, 2000

[5] Y. Benjamini & D. Yekutieli *The Control of the False Discovery Rate in Multiple Testing under Dependency.* The Annals of Statistics, Vol. 29, No. 4, pp. 1165-1188, 2001

[6] S. Holm *A Simple Sequentially Rejective Multiple Test Procedure*, Scandinavian Journal of Statistics, Vol. 6, No. 2, pp. 65-70, 1979

[7] Y. Benjamini & D. Yekutieli & A.Krieger *Adaptive Linear Step-up Procedures That Control the False Discovery Rate.* Biometrika, Vol. 93, No. 3, pp. 491-507, 2006

[8] B. Efron *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, Institute of Mathematical Statistics Monographs. Cambridge University Press, 2010

[9] B. Efron *Doing Thousands of Hypothesis Tests at the Same Time*, Metron - International Journal of Statistics Vol. 65, No. 1, pp. 2-31, 2007

[10] D.Singh & al. *Gene expression correlates of clinical prostate cancer behavior*, Cancer Cell, Vol. 1, Issue 2, pp. 203-209, 2002.

[11] D.Cox & D.Hinkley *Theoretical Statistics* Chapman and Hall, 1974

[12] J. Kalbfleisch *Probability and Statistical Inference. Vol. 2: Statistical Inference* Springer-Verlag,1985

[13] M. Glickman & S.Rao & M. Schultz *False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies* Journal of Clinical Epidemiology Vol 67, Issue 8, pp. 850-857, 2014

Liitteet

Lyhenneluettelo

Suure	Määritelmä
FDR	hylkäysvirheaste
α	merkitsevyystaso
H_0	nollahypoteesi
H_1	vastahypoteesi
FWER	yhdistetty merkitsevyystaso
p	p-arvo
$\overset{Bayes}{Fdr}$ ja $\Phi(\mathcal{Z})$	Bayesiläinen hylkäysvirhesaste
$\overline{Fdr}(\mathcal{Z})$, $\overline{Fdr}(z)$ ja $\overline{Fdr}(z_0)$	empiirinen estimaatti $\overset{Bayes}{Fdr}$:lle
$N_0(z_0)$, $N_0(z_1)$ ja $N(z_0)$	nollageenien, ei-nollageenien ja hylättyjen nollahypoteesien lukumäärä (kts. kuva (7))
Fdp ja Q	virheellisten löydösten osuus
BH-menetelmä	Benjaminin & Hochbergin kehittämä valintamenetelmä
BKY-menetelmä	Benjaminin & Hochbergin & Kriegerin kehittämä adaptiividen valintamenetelmä
LSL-menetelmä	Alimman tason käyrään perustuva menetelmä

R-koodit

```
# Pro gradu Ida Isaksson

## Koodien pohjana on kaytetty viitetta
##Mike Love MIT PH525x series - Biomedical Data
## Science course

#Asetetaan tarvittavat kirjastot ja haetaan data
library(sda)
library(locfdr)
library(fdrtool)
library(MASS)
library(zoom)
data(singh2002)
setwd("C:/Users/idais/Documents/Gradu")
Xtrain = singh2002$x
control<-prostmat[1:50]
test<-prostmat[51:102]

load("prostz.RData")
zval<-prostz


#histogrammi ja normaalikayra
h<-hist(zval, breaks=84,main="",xlab="Z-arvo",ylab="Frekvenssi")
xfit <- seq(min(zval), max(zval), length = 84)
yfit <- dnorm(xfit, mean = 0, sd = 1)
yfit <- yfit * diff(h$mids[1:2]) * length(zval)
lines(xfit, yfit, col = "black", lwd = 2)
legend(-3.9,200,"N(0,1)", cex=0.8,box.lty=0,lty=1,lwd=2)


#suoritetaan N=6033 parittaista t-testia ei muunnetulle datalle)
control<-prostmat[1:50]
test<-prostmat[51:102]
p = sapply(1:length(prostmat[,1]), function(i)
t.test(control[i,],test[i,])$p.val)

# haetaan p-arvot
sum(p<0.05)
sum(p.adjust(p, "BH")<0.05)
```

```
# BH:n valintamenetelma
q <- 0.1
i = seq(along=zval)
N <- 6033
mypar(1,2)
plot(i,sort(p))
plot(i[1:100],sort(p)[1:100],main="",ylab="p-arvo",xlab="i")
abline(0,i/N*q)


#lasketaan suurin indeksi i jolle FDR-maaritelma patee
k <- max( which( sort(p) < i/N*q) )
cutoff <- sort(p)[k]
cat("k␣=",k,"p-value␣cutoff=",cutoff)


fdr = fdrtool(p, statistic="pvalue")
fdr$qval # estimoidut Fdr -arvot
fdr$lfdr # estimoidut lfdr-arvot


#Haetaan z-arvot datalle
zval
N<-6033
sum(abs(zval)>=1.96)
sum(abs(zval)>=abs(qnorm(0.05/6033)))
N_0<-sum(zval<=-3)
Fdr_est <- N*pnorm(-3)/sum(zval<=-3)
sum(p<=0.05)



#BAYES FDR KONTROLLOINTI

#Vasemman puoleiset z-arvot
zvall<-sort(zval[zval<0])
il = seq(along=sort(zvall))
#Oikean puoleiset z-avrot
zvalr<-sort(zval[zval>0],decreasing = T)
ir = seq(along=zvalr)
q<-0.1
N=6033
```

```
#etsitaan suurin indeksi k, jolle ehto F(zi)/ (i/N) < q patee
(vasen)
kl <- max( which(pnorm(zvall[il])/(il/N) < q ))
cutoff <- zvall[kl]
cat("k␣=",kl,"z-value␣cutoff=",cutoff)

#etsitaan suurin indeksi k, jolle ehto F(zi)/ (i/N) < q patee
(oikea)
kr <- max(which((1-pnorm(zvalr[ir]))/(ir/N) < q ) )
cutoff <- zvalr[kr]
cat("k␣=",kr,"z-value␣cutoff=",cutoff)



#LOCAL FALSE DISCOVERY RATE z.arvoille

a<-hist(zval,breaks=89, main="",xlab = "z-arvo",ylab="Frekvenssi")

est<-locfdr(zval, bre = 89, df =7, pct = 0, pct0 = 1/4,
 nulltype = 0, type =
              1, plot = 1, main = "␣", sw = 0)

br<-a$breaks
xk<-a$mids[80]
yk<-a$counts[80]
which(a$counts==4,a$counts)


# local fdr < 0.2
est$z.2


#Lasketaan estimaatti binille K=90
d<-0.1
pi<-1
N<-6033
e0<-N*d*pi*dnorm(xk)
e0/yk

mat <- est $mat
#ne arvot, jotka toteuttavat ehdon fdr <=0.2
length(est$fdr[est$fdr<=0.2])
```

```
#vasemmanpuoleiset z-arvot <= 0.2
zval[zval<=-3.415117]
#oikeapuoleiset z-arvot <= 0.2
zval[zval>=3.371268]



# Osajoukkojen keskipiste x_{k}
bins<-mat[,1]

#haetaan vasemman ja oikean puoleiset keskipisteet
left00<-mat[mat[,1]<=0,]
rigth00<-mat[mat[,1]>=0,]

#vasemman ja oikeanpuoleiset Fdr -arvot
left<-left00[,3]
rigth<-rigth00[,4]


#piirretaan kuva
h<-mat[mat[,2]!=1,]
h[,2]
plot(mat[,2])
plot(left,type="l")
plot(rigth ,type="l",add=T)
fdr<-mat[,8]

plot(bins[bins<=0],left,type="l",xlim=c(-4, 4),ylim=c(0, 1),lty=3,
ylab="Fdr",xlab="z-arvo",main="") lines(bins[bins>=0],rigth,lty=4)
  lines(bins,fdr,type="l") abline(v=0,lty=3)
  legend(-4.3, 1, legend=c("lokaali fdr",
"Vasemmanpuoleinen Fdr","Oikeanpuoleinen Fdr"),
        lty=c(1,3,4), cex=0.8,box.lty=0)




# Pro gradu Ida Isaksson  SIMULAATIO
```

```r
#simuloidaan N:n kokoinen ryhmat populaatiosta N(3,1)
set.seed(4)
pop = rnorm(1000,3)
N <- 10
m <- 10000


# Suoritetaan 10 000 t-testia
set.seed(1)
pvals <- replicate(m,{
  control <- sample(pop,N)
  treatment<- sample(pop,N)
  t.test(treatment,control)$p.value
})

sum(pvals < 0.05)


#Asetetaan 90% nollahypoteeseista tosiksi
N <- 10
m <- 10000
p_0<-0.9
alpha<-0.05
m_0<-m*p_0
m_1<-m-m_0
nollahypoteesit<-c( rep(TRUE,m_0), rep(FALSE,m_1))



#suoritetaan jalleen 10000 t-testia ja taulukoidaan tulokset
set.seed(1)
calls <- sapply(1:m, function(i){
  control <- sample(pop,N)
  treatment <- sample(pop,N)
  if(!nollahypoteesit[i]) treatment <- treatment+1
  ifelse( t.test(treatment,control)$p.value < alpha,
          "Merkitseva",
          "Ei-merkitseva")
})
```

```
nolla_hypo <- factor(nollahypoteesit, levels=c("TRUE","FALSE"))
table(nolla_hypo ,calls)
```

```
library(genefilter)
set.seed(1)
#ROWTTEST paketti suorittaa samanaikaisia
## t-testeja simulaatiota varten
g <- factor( c(rep(0,N),rep(1,N)) )
#Simulaatioiden lkm
B <- 1000
Qs <- replicate(B,{
  #"Kontrolli" rivit testeja, sarakkeet "yksiloita"
  controls <- matrix(sample(pop, N*m, replace=TRUE),nrow=m)

  #"Vaste" rivit testeja, sarakkeet "yksiloita"
  treatments <-  matrix(sample(pop, N*m, replace=TRUE),nrow=m)

  #Vaihdetaan vaste -ryhmasta 10% epatosiksi nollahypoteeseiksi
  a<-length(treatments[which(!nollahypoteesit),])
  treatments[which(!nollahypoteesit),]
  <- treatments[which(!nollahypoteesit),]+1
  #Yhdistetaan tulokset
  dat <- cbind(controls,treatments)

  calls <- rowttests(dat,g)$p.value < alpha
  R=sum(calls)
  Q=ifelse(R>0,sum(nollahypoteesit & calls)/R,0)
  return(Q)
})


library(rafalib)
mypar(1,1)
# Q:n jakauma
hist(Qs,main="",xlab="Q",ylab="Frekvenssi")


FDR=mean(Qs)
print(FDR)
```

```
#Suoritetaan edellinen uudelleen, kayttaen BH:n valintamenetelmaa
set.seed(1)
controls <- matrix(sample(pop, N*m, replace=TRUE),nrow=m)
set.seed(2)
treatments <-  matrix(sample(pop, N*m, replace=TRUE),nrow=m)
treatments[which(!nollahypoteesit),]
<-treatments[which(!nollahypoteesit),]+1
dat <- cbind(controls,treatments)
pvals <- rowttests(dat,g)$p.value


q <- 0.1
i = seq(along=pvals)

dev.off()
plot(i,sort(pvals))
abline(0,i/m*q)
#Naytetaan 120 pieneninta p-arvoa
plot(i[1:120],sort(pvals)[1:120],main="",ylab="p-arvo",xlab="i")
abline(0,i/m*q)
abline(alpha/m,0,lty=2)
legend(2.5,0.0012, legend=c("BH-metodi", "Bonferroni-korjaus"),
       lty=c(1,2), cex=0.8,box.lty=0)

#lasketaan suurin indeksi i jolle FDR-maaritelma patee
k <- max( which( sort(pvals) < i/m*q) )
cutoff <- sort(pvals)[k]
cat("k␣=",k,"p-value␣cutoff=",cutoff)


fdr <- p.adjust(pvals, method="fdr")
fdr[fdr<=0.1]


#Bonferroni hylkaykset
sum(pvals<=0.05/m)



#Suoritetaan 1000 simulaatiota
```

```
alpha <- 0.1
B <- 1000
set.seed(2)
res <- replicate(B,{
  controls <- matrix(sample(pop, N*m, replace=TRUE),nrow=m)
  treatments <-  matrix(sample(pop, N*m, replace=TRUE),nrow=m)
  treatments[which(!nollahypoteesit),]
  <-treatments[which(!nollahypoteesit),]+1
  dat <- cbind(controls,treatments)
  pvals <- rowttests(dat,g)$p.value
#FDR valinta
  calls <- p.adjust(pvals,method="fdr") < alpha
  R=sum(calls)
  Q=ifelse(R>0,sum(nollahypoteesit & calls)/R,0)
  return(c(R,Q))
})
Qs <- res[2,]
mypar(1,1)
hist(Qs,main="",xlab="Q",ylab="Frekvenssi")
mean(Qs)


FDR=mean(Qs)
print(FDR)
```