Lauri Nummenmaa

Efektikoko psykologisessa tutkimuksessa

Psykologisessa tutkimuksessa tilastollinen päätöksenteko perustuu tyypillisesti Fisherin määrittelemiin alfatasoihin ja nollahypoteesin merkitsevyystestaamiseen (NHST). Tällainen menettely ei kuitenkaan ilmaise mitään tutkimuksessa muuttujien välisen yhteyden voimakkuudesta. Lisäksi NHST:n tulosten tulkinta on ongelmallista. Käsittelen tässä katsauksessa yleisimpiä ongelmia ja päättelyvirheitä, jotka liittyvät NHST:n soveltamiseen tutkimusaineiston analyysissa. Vertailen NHST:n vaihtoehdoksi ja täydennykseksi esitettyjä efektikoon arvioimiseen perustuvia tunnuslukuja (d, g, Δ, r, Φ, η ja ω) ja arvioin niiden soveltuvuutta psykologisen tutkimuksen tulosten esittämiseen. Esitän myös suosituksen efektikoon estimaattien käyttämisestä tutkimustuloksia raportoitaessa.

Psykologinen tutkimus perustuu numeerisessa muodossa olevan tutkimusaineiston analysoimiseen tilastollisten menetelmien avulla. Tilastollisten menetelmien tavoitteena on tyypillisesti tarkastella aineistossa olevaa signaali/kohina -suhdetta (Killeen, 2005). Signaalilla viitataan tutkittavien muuttujien väliseen mahdolliseen assosiaatioon ja kohinalla otanta- ja mittausvirheestä johtuvaan satunnaisvaihteluun. Tilastollisten menetelmien avulla on siten mahdollista arvioida, johtuuko kahden muuttujan havaittu assosiaatio todellakin muuttujien välisestä yhteydestä vai onko se aiheutunut sattumalta. Psykologiassa sovellettavat tilastolliset analyysimenetelmät ovat kehittyneet viimeisten kahdenkymmenen vuoden aikana huomattavasti. Esimerkiksi Bayesilaisten menetelmien käyttäytymistieteelliset sovellukset (Gill, 2002), latentit kasvukäyrämallit (Muthèn & Muthèn, 2000) ja riippumattomien komponenttien analyysi (Stone, 2002) ovat mahdollistaneet tutkimusaineistojen entistä tarkemman ja monipuolisemman kuvailemisen sekä analysoimisen.

Psykologiassa käytettävät tilastollisen päättelyn periaatteet eivät ole kuitenkaan kehittyneet yhtä nopeasti. Yleisimmin ilmiöitä koskevaan tilastolliseen päättelyyn sovelletaan edelleen kiistanalaista ja monessa suhteessa ongelmallista menettelyä, joka tunnetaan nimellä nollahypoteesin merkitsevyystestaus (Null Hypothesis Significance Testing, NHST). NHST:lle on esitetty lukuisia vaihtoehtoisia ja täydentäviä menettelyitä, joiden avulla tutkimusaineistoa koskeva päättely voidaan suorittaa tarkemmin ja paremmin. Nämä menettelyt eivät kuitenkaan ole yleistyneet kuin vasta viime vuosien aikana. Käsittelen tässä katsauksessa keskeisempiä NHST-menettelyyn liittyviä ongelmia ja arvioin sille esitettyjä vaihtoehtoisia ja täydentäviä menetelmiä, efektikoon estimaatteja.

NHST JA SIIHEN LIITTYVÄT ONGELMAT

Fisher (Fisher & Bennett, 1925 / 1990) esitti nollahypoteesin merkitsevyystestaamisen periaatteen

sellaisena kuin se nykyään tunnetaan. Jos koeasetelmassa on *j* solua, niin kontrasti ψ määritellään painokertoimien c_j avulla seuraavasti:

$$\psi = c_1\mu_1 + c_2\mu_2 \ldots + c_j\mu_j \quad (1.1)$$

missä

$$c_1 + c_2 + \ldots c_j = 0$$

Testin *p*-arvo määritellään tällöin laskemalla määrätylle kontrastille ψ, kuinka todennäköistä on saada tilastollinen tunnusluku joka on suurempi kuin $|\psi|$ välillä $[\,|\psi|,\infty]$. Tämä siis ilmaisee todennäköisyyden $P(x \geq \psi \mid H_0)$. Merkitään dataa *D*:llä, jolloin yleisessä tapauksessa tarkastellaan todennäköisyyttä $P(D \mid H_0)$. NHST ilmoittaa siten todennäköisyyden sille, että otannan avulla saadaan havaitun kaltainen data, jos nollahypoteesi pitää paikkansa. Tällaisessa muodossa esitetyn NHST:n soveltamiseen liittyy kuitenkin neljä suurta ongelmaa, joita käsittelen lyhyesti ennen kuin siirryn tarkastelemaan NHST:lle vaihtoehtoisia lähestymistapoja. Ongelmat ovat

1. NHST:ssa testataan todennäköisyyttä $P(D \mid H_0)$, eikä tutkijan kannalta kiinnostavampaa todennäköisyyttä $P(H_0 \mid D)$ (Cohen, 1994).
2. Yleisimmin testattava nollahypoteesi (H_0: $\mu_1 - \mu_2 = 0$) ei pidä koskaan paikkaansa (Tukey, 1991).
3. NHST ei kvantifioi havaitun ilmiön amplitudia ja on siten suboptimaalinen menettely teorianmuodostuksessa (Loftus, 1996).
4. Alfatason valinta on arbitraarinen (Glass, McGaw, & Smith, 1981).

$P(D \mid H_0) \neq P(H_0 \mid D)$

Fisherin tapa nollahypoteesien testaamiseen on intuitiivisesti mielekäs. NHST:n tulosten tulkitseminen on kuitenkin ongelmallista. NHST:ssa lasketaan $P(D \mid H_0)$, eli todennäköisyys sille, että data havaitaan sillä ehdolla, että nollahypoteesi on asetettu oikein. Tämä ei kuitenkaan ole yleensä tutkimuksen kannalta mielenkiintoista – tutkijaa kiinnostaa paljon useammin selvittää, mikä on $P(H_0 \mid D)$, eli mikä on todennäköisyys sille, että nollahypoteesi on voimassa, jos data on havaitun kaltainen. On huomattava, että intuition vastaisesti $P(D \mid H_0) \neq P(H_0 \mid D)$. Jotta voisimme NHST:n tulosten perusteella laskea posteriorisen todennäköisyyden $P(H_0 \mid D)$, meidän tulee käyttää Bayesin teoreemaa (Bayes, 1764). Bayesin teoreeman mukaan

$$P(B \mid A) = \frac{P(A \cap B)}{P(B)} \quad (1.2)$$

mikä siis NHST:n tapauksessa tarkoittaa

$$P(H_0 \mid D) = \frac{P(D \cap H_0)}{P(H_0)}$$

Jotta voisimme laskea posteriorisen todennäköisyyden $P(H_0 \mid D)$, meidän tulisi tietää priorinen todennäköisyys $P(H_0)$, eli ennen tutkimusta tiedossa ollut todennäköisyys sille, että nollahypoteesi pitää paikkansa. Yleensä tämä ei kuitenkaan ole tiedossa, muutenhan NHST:n tekeminen ei olisi lainkaan tarpeellista. Bayesilaisessa päättelyssä (ks. esim. Gill, 2002) ongelma ratkaistaan siten, että priorinen todennäköisyys (tai sen jakauma) asetetaan kaiken mahdollisen käytettävissä olevan prioorisen tiedon perusteella ja testisuureen posteriorinen jakauma määritellään tämän perusteella. Tätä tietoa voidaan taas vastaavasti käyttää määriteltäessä prioreja seuraavassa tutkimuksessa.

Milloin nollahypoteesi voi olla oikein asetettu?

Jos emme kuitenkaan [syystä tai toisesta] halua siirtyä Bayesilaiseen tilastolliseen päättelyyn, Cohen (1994) suosittelee että NHST:ta käytettäisiin ainoastaan sellaisessa ”vahvassa” muodossa kuin Popper (1959) on esittänyt. Tällöin tieteellisen teorian tulee edetä yrityksinä kumota olemassa olevia teorioita, mikä onkin mahdollista NHST:n avulla. Sen sijaan NHST:n avulla ei ole mahdollista todistaa teorioita oikeaksi hylkäämällä nollahypoteeseja. Tämä on ilmeistä jos ajatellaan, millaisia kontrasteja ψ testattaessa yleisimmin käytetyt nolla- (H_0: $\mu_1 - \mu_2 = 0$) ja vaihtoehtoinen hypoteesi (H_1: $\mu_1 - \mu_2 \neq 0$) tyypillisesti ovat. Tällä tavoin määritelty nollahypoteesi on triviaalisti epätosi ja vaihtoehtoinen hypoteesi triviaalisti tosi. Lähestulkoon minkä tahansa kahden jakauman odotusarvoissa havaitaan todennäköisesti eroa, jos mittaustarkkuus on riittävän suuri. Tällaisen nollahypoteesin voidaan osoittaa olevan *aina* väärin asetettu. Määritellään $f(x)$=normaalijakauman kertymä-

funktio ja $g(x)$=t-jakauman kertymäfunktio. Tällöin

$$\lim_{x\to\infty} f(x) = 1$$
$$\lim_{x\to\infty} g(x) = 1 \quad (1.3)$$

Tämä siis tarkoittaa, että jos kaikki muut tekijät pysyvät vakioina, niin em. kertymäfunktioiden arvo lähestyy ykköstä kun otoskoko lähestyy ääretöntä. Tarpeeksi suurella otoskoolla siis *mikä tahansa* keskiarvojen ero on tilastollisesti merkitsevä. Käytettäessä NHST:ta tällä tavoin asetetun nollahypoteesin kumoamiseen voidaan itse asiassa ainoastaan osoittaa, että käytetty tutkimusasetelma oli riittävän vahva havaitsemaan olemassa olevan keskiarvojen eron (Kirk, 1996).

NHST ei kvantifioi tutkimuksessa havaitun ilmiön voimakkuutta

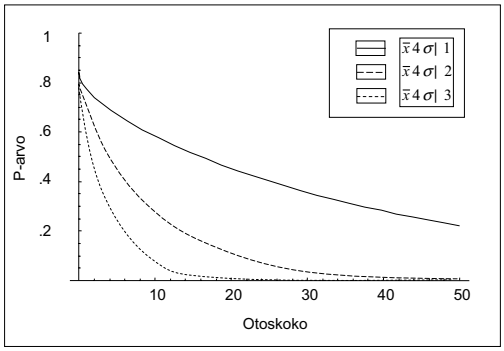
Useimmissa käytetyistä tilastollisissa testeissä *p*-arvo riippuu sekä signaali/kohina -suhteesta että otoskoosta. Intuitiivisesti voidaan ajatella, että NHST:ssa *p*-arvot muodostuvat seuraavasti (Nummenmaa, 2004):

$$\text{p}-\text{arvo} = \frac{1}{\text{efekti}\times\text{otoskoko}} \quad (1.4)$$

Tästä siis seuraa, että suurissa otoksissa pienetkin efektit ovat tilastollisesti merkitseviä ja pienissä otoksissa efektin on oltava suuri, jotta se olisi tilastollisesti merkitsevä. Tarkastellaan esimerkkinä yhden otoksen Z-testiä

$$Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \quad (1.5)$$

Testisuure Z noudattaa normaalijakaumaa parametrein [0, 1]. Sovitaan, että. Tarkastellaan nyt, miten testin *p*-arvo muuttuu otoskoon funktiona kun $\bar{x}-\mu=1$ (1), $\bar{x}-\mu=2$ (2) ja $\bar{x}-\mu=3$ (3) (Kuva 1). Kun keskiarvojen erotus on pieni (tässä 1), ei *p*-arvosta tule tilastollisesti merkitsevää (<.05) edes viidenkymmenen henkilön otoksella. Kun taas keskiarvojen erotus on suuri (tässä 3), niin keskiarvojen erotuksesta tulee tilastollisesti merkitsevä jo alle kahdenkymmenen hengen otoksella.



Kuva 1. NHST:n p-arvon muuttuminen otoskoon funktiona.

P-arvot ovat siis riippuvaisia sekä efektin koosta että otoskoosta, mutta *p*-arvoissa nämä kaksi tietoa tiivistetään yhteen tunnuslukuun. Koska tunnusluvun suuruus riippuu efektin suuruuden lisäksi myös otoskoosta, NHST:ta käytettäessä on siten suuri riski hyväksyä efekti, jonka voimakkuus on triviaali (Chow, 1988). Kääntäen on myös mahdollista hylätä voimakkuudeltaan suuri efekti riittämättömän otoskoon takia (Kirk, 1996). Koska yksittäisen tutkimuksen otoskoko on mielivaltainen, pelkän *p*-arvon raportoiminen ei siis riitä.

Alfatasot ovat arbitraarisia

Eräs useimmin NHST:ta kohtaan esitetyistä kritiikeistä on alfatasojen arbitraarisuus (Kirk, 1996). NHST:ta käytettäessä ilmiö ikään kuin muuttuu todeksi, kun *p*-arvo alittaa sovitun kriittisen rajan – tyypillisesti .05:n, mutta rajan asettamiselle ei ole objektiivista perustelua. On siis mahdollista, että kaksi tutkijaa voi havaita samanlaisissa koeasetelmissa täsmälleen saman keskiarvojen erotuksen, mutta erilaisista otoskoista johtuen he saavat NHST:n tuloksena *p*-arvot .07 ja .05. NHST:a käytettäessä toinen tutkijoista siis saa tutkimuksellaan tukea nolla- ja toinen vaihtoehtoiselle hypoteesille. Mutta, jos kummatkin tutkijat olisivatkin valinneet alfatasoksi .01:n, niin molemmat olisivat hylänneet vaihtoehtoisen hypoteesin. Koska alfatasoihin perustuvan päättelyn lopputulos riippuu sekä otoskoosta että valitusta alfatasosta, Rosenthal ja Rubin (1989) tiivistävätkin alfatasojen mielekkyyden klassisessa Psychological Bulletin -ar-

tikkelissaan lauseeseen ”Surely, God loves the .06 almost as much as the .05”.

EFEKTIKOKO

NHST:a käytettäessä muodostetaan binäärinen (tosi/epätosi) väite tarkastellusta ilmiöstä. Jotta useissa tutkimuksissa kerättyä tietoa voitaisiin yhdistää, olisi tarkoituksenmukaista että jokainen tutkimus tuottaisi – mielellään vähintään välimatka-asteikollisen – kvantifikaation tarkasteltujen muuttujien välisestä yhteydestä. Tutkimustulosten arvioimiseen tarvitaan siis sellaisia tunnuslukuja, joiden avulla (i) voidaan tarkastella mitattujen muuttujien välisen yhteyden voimakkuutta ja (ii) voidaan yhdistää tuloksia yli tutkimusten. Tilastotieteilijät ovat jo vuosia suositelleet, että NHST:n lisäksi tutkimusraporteissa tulisi ilmoittaa jokin havaitun efektin voimakkuutta paremmin kuvaava luku, esimerkiksi *MSE* tai *efektikoko* (Chow, 1988). Kuitenkin vasta viides painos Publication Manual of the American Psychological Association:sta (American Psychological Association, 2001) sisältää eksplisiittisen vaatimuksen efektikokojen raportoimisesta.

Efektikoko voidaan määritellä usealla tavalla, mutta tyypillisesti se on otoskoosta riippumaton numeerinen estimaatti, joka kvantifioi ψ:n eli riippumattoman muuttujan riippuvassa muuttujassa aiheuttaman vaihtelun. Yleinen käytäntö on, että negatiiviset arvot viittaavat efektiin, joka on vaihtoehtoisen hypoteesin vastainen ja positiiviset sellaiseen, joka on vaihtoehtoisen hypoteesin suuntainen. Yleisimmin käytetyt efektikokostatistikat perustuvat *d*:hen (Cohen, 1969), *r*:ään sekä Φ:hin (Rosenthal, 1984; Rosenthal & Rubin, 2003), sekä η:aan ja ω:aan (Olejnik & Algina, 2003). Nämä voidaan jakaa kahteen kategoriaan: standardoituun keskiarvojen erotukseen (*d*) ja riippuvan ja riippumattoman muuttujan väliseen assosiaatioon perustuviin (r, Φ, η, ω) tunnuslukuihin.

Ryhmäkeskiarvoihin ja -keskihajontoihin perustuvat efektikoon estimaatit: *d*, *g* ja Δ

Tarkastellaan aluksi *d*:tä ja siihen perustuvia efektikoon estimaatteja. Cohenin *d* (signaalinkäsittelyssä käytetään merkintää d') ilmaisee efektin suuruuden ryhmäkeskiarvojen standardoituna erotuksena. Yleisessä muodossa *d*-tyyppinen efektikoko ilmaistaan seuraavasti:

$$\text{(1.6)} \quad \text{efekti} = \frac{\psi}{s}$$

missä ψ on tarkasteltava kontrasti ja s on hajontatermi, johon kontrastia verrataan. Riippuen koeasetelmasta s voidaan määritellä useammallakin tavalla, ja jokainen määrittelyistä johtaa erilaiseen efektikoon estimaattiin. Näistä yleisimmin käytetyt (Olejnik & Algina, 2000) ovat (i) vertailtavien solujen yhdistetty keskihajonta s_p (Cohenin *d*), (ii) kaikkien asetelman solujen yhdistetty keskihajonta S_p (Hedgesin *g*) ja kontrolliryhmän keskihajonta S_c (Glassin Δ):

$$\text{(1.7)} \quad \text{Cohenin } d = \frac{\bar{x}_1 - \bar{x}_2}{s_p}$$

$$\text{Hegdesin } g = \frac{\bar{x}_1 - \bar{x}_2}{S_p}$$

$$\text{Glassin } \Delta = \frac{\bar{x}_1 - \bar{x}_2}{S_c}$$

Jos koe- ja kontrolliryhmien varianssien yhtäsuuruusoletus on voimassa, sekä *d*, *g* että Δ ovat yhtä suuria. Jos oletus ei ole voimassa, tutkijan tulee valita sellainen hajonnan estimaatti, joka kuvaa kontrastin tarkoituksenmukaista mittayksikköä käyttäen. Jos käytetään kontrolliryhmän keskihajontaa (Δ) tai kaikkien solujen yhdistettyä keskihajontaa (g), on efektikoon nimittäjään tuleva hajontatermi sama kaikissa mahdollisissa kontrasteissa ja kontrasteille lasketut efektikoot ovat keskenään vertailukelpoisia. Haittapuolena kuitenkin on se, että varianssitermi ei välttämättä kuvaa tarkasteltavan kontrastin hajontaa kovinkaan hyvin. Jos taas hajontatermi lasketaan kontrastikohtaisesti, kuvaa hajontatermi paremmin jokaista yksittäistä kontrastia, mutta eri kontrasteille lasketut efektikoot eivät olekaan enää vertailukelpoisia keskenään.

Cohenin *d*:hen perustuvien estimaattien eräs ongelma on, että periaatteessa ne voivat vaihdella välillä $[-\infty, \infty]$. Tyypillisissä tutkimuksissa *d*-perustaisten estimaattien vaihteluväli rajoittuu kuitenkin noin [-1, 1]:een. Kuinka suuri *d*-perustaisen efektin sitten tulisi olla, jotta efektiä voitaisiin pitää ”riittävän” suurena? Tähän ei voida antaa mi-

tään yksikäsitteistä vastausta. Efektikoon tilastollista merkitsevyyttä ei tietenkään kannata ruveta testaamaan, silloinhan täytyisi laskea taas toisen kertaluvun efektikoko ”efektikoon *p*-arvolle” ja niin edelleen. Yleensä käytetäänkin Cohenin (1992) esittämiä suuntaa antavia arvoja. Tällöin keskinkertaisen efektin suuruus on .5. Tämänsuuruinen efekti näkyy jo aineistoa silmäilemällä. Pieni efekti on suuruudeltaan .2 ja suuri vastaavasti .8. Näitä suuntaa-antavia arvoja ei kuitenkaan tule tulkita samaan tapaan kuin p-arvon kriittisiä rajoja, vaan ne ovat ainoastaan efektin suuruuden tulkinnassa avuksi käytettäviä arvoja.

Korrelaatiokertoimeen perustuva efektikoon estimointi

Cohenin *d* soveltuu käytettäväksi sellaisissa tutkimusasetelmissa, joissa on tarkasteltu vain asetelman kahden solun välistä kontrastia $\mu_1 - \mu_2$. Tämä onkin tyypillisin kontrasti. Aina efektiä ei kuitenkaan määritellä kahden solukeskiarvon välisenä erotuksena – esimerkiksi tarkasteltaessa kahden jatkuvan muuttujan välistä assosiaatiota, Cohenin *d* ei sovellu käytettäväksi efektikoon estimaattina. Tämän vuoksi Rosenthal ja Rubin ovatkin ehdottaneet, että efekti määriteltäisiin riippuvan ja riippumattoman muuttujan yhteisenä vaihteluna (Rosenthal, 1984; Rosenthal & Rubin, 2003). Tällöin efektikoko määritellään muuttujien *i* ja *j* välisen korrelaatiokertoimena *r*. Jos muuttujat *i* ja *j* ovat numeerisia, *r* määritellään yksinkertaisesti muuttujien tyypistä riippuen Pearsonin tulomomenttikertoimena tai Spearmanin järjestyskorrelaatiokertoimena. Jos toinen muuttuja on kaksiluokkainen ja toinen jatkuva, käytetään piste-biseriaalista korrelaatiokerrointa. Voidaan osoittaa (Rosenthal & DiMatteo, 2001), että *r* voidaan laskea myös jälkikäteen *t*-arvoista, *F*-arvoista, joiden vapausasteiden osoittajassa on 1 sekä yhden vapausasteen χ^2-arvoista:

$$(1.8) \qquad r = \sqrt{\frac{t^2}{t^2 + df}}$$

$$(1.9) \qquad r = \sqrt{\frac{F}{F + df}}$$

$$(1.10) \quad r = \sqrt{\frac{\chi^2}{N}}$$

Lisäksi *r* voidaan estimoida pelkän *p*-arvon avulla, jos muuta informaatiota ei ole käytettävissä. Tällöin määritellään *p*-arvoa vastaava yksisuuntaisen testin *Z*-arvo ja lasketaan efektikoko seuraavasti:

$$(1.11) \quad r = \frac{Z}{\sqrt{N}}$$

Jos vertailtavat ryhmät ovat riippumattomia, *r* on yksinkertaisesti riippuvan muuttujan ja dummy-koodatun lohkotekijän välinen piste-biseriaalinen korrelaatio. Jos taas ryhmät eivät ole toisistaan riippumattomia, *r* on ryhmään kuulumisen ja riippuvan muuttujan välinen osittaiskorrelaatio, josta on ositettu toistettujen mittausten indikaattorimuuttujan vaikutus. *R*-tyyppinen efektikoko voidaan määrittää myös silloin, jos tarkasteltavat muuttujat ovat kategorisia. Jos tarkastellaan efektin suuruutta 2x 2 -kontingenssitaulussa, niin Cramerin *Φ*-kerrointa $\left(\sqrt{\frac{\chi^2}{n}}\right)$ voidaan käyttää *r*-tyyppisenä efektikoon estimaattina. Mielivaltaisessa *m* x *n* –kontingenssitaulussa puolestaan kontrasti voidaan laskea seuraavasti (Rosenthal & Rosnow, 1991, s. 538)

$$(1.12) \quad Z = \frac{\sum P \times c}{\sqrt{\sum \frac{P(1-P)}{N} \times c^2}}$$

missä *c* on kontrastin painokerroin, *P* on sarakkeen vertailtavien solujen [tai niiden yhdistelmien] frekvenssien suhde ja *N* sarakkeen kokonaisfrekvenssi.

Efektikoon estimoiminen *r*:n avulla tarjoaa useita etuja *d*:hen perustuviin estimaatteihin verrattuna. Suurin näistä lienee se, että *r* voidaan laskea useammanlaisissa asetelmissa kuin *d*. Toinen huomattava etu on, että korrelaatioon perustuvana efektikoon estimaattina *r* on standardoitu, se vaih-

telee aina välillä [-1,1]. Tällöin kaikkien mahdollisten *r*-tyyppisten efektikokojen yhdistäminen ja vertaileminen on periaatteessa yksinkertaista. Useissa tapauksissa *r*:n tulkinta on myös suoraviivaisempaa – siinä missä *d*:n, *g*:n ja Λ:n suuruus riippuu mitattujen muuttujien hajonnoista, *r* on yksinkertainen suhdeluku, joka ilmoittaa, kuinka paljon yhteistä vaihtelua riippuvalla ja riippumattomalla muuttujalla on. Efektikoot ovat kuitenkin vaihdannaisia – vastaavissa asetelmissa lasketut *r* ja *d* voidaan tarvittaessa jälkeenpäin muuntaa toisikseen:

$$(1.13) \quad r = \sqrt{\frac{d^2}{d^2 + 4}}$$

$$(1.14) \quad d = \frac{2r}{\sqrt{1 - r^2}}$$

Samaan tapaan kuin *d*:hen perustuvien estimaattien, niin myös *r*:n suuruuden tulkintaan voidaan antaa suuntaviivoja. Käyttämällä kaavaa 1.14 saadaan Cohenin (1988) *d*:n suuruuksia vastaaviksi *r*:n rajoiksi 0.1 (pieni efekti), 0.24 (keskisuuri efekti) ja 0.37 (suuri efekti).

Efektikoon η- ja ω-estimaatit

Efektikoon η- ja ω-estimaatit ovat läheistä sukua *r*-estimaateille. Toisin kuin *r*, η- ja ω-estimaatit ilmoittavat efektin suuruuden selitetyn varianssin ja kokonaisvarianssin osamääränä (Olejnik & Algina, 2003). *R*-estimaattien avulla voidaan tarkastella pääasiassa yksinkertaisia kontrasteja, mutta η- ja ω -estimaattien avulla voidaan arvioida myös ANOVA:n päävaikutusten efektikoot. Koska efekti määritellään nyt selitetyn varianssin osuutena kokonaisvarianssista, tulee myös käytetty tutkimusasetelma huomioida efektikokoa laskettaessa. Tarkastellaan aluksi koeasetelmaa, jossa kaikki tekijät ovat täysin satunnaistettuja. Tällöin efektikoon estimaattina käytettävä $\hat{\eta}^2$ määritellään yksinkertaisesti tarkasteltavan kontrastin tai päävaikutuksen neliösumman SS_{effect} ja yhteisneliösumman SS_{total} osamääränä:

$$(1.15) \quad \hat{\eta}^2 = \frac{SS_{effect}}{SS_{total}}$$

Tutkimuksessa tarkasteltavien tekijöiden lukumäärä vaikuttaa kuitenkin tällä tavoin määritellyn $\hat{\eta}^2$:n suuruuteen, koska jokaisen tekijän aiheuttama vaihtelu vaikuttaa yhteisneliösummaan SS_{total} (Olejnik & Algina, 2003). Tällöin täsmälleen samaa ilmiötä tarkastelevissa tutkimuksissa pelkästään tekijöiden lukumäärä saattaa aiheuttaa ei-satunnaista vaihtelua $\hat{\eta}^2$:n estimaatteihin. Tämän vuoksi efektikoon estimaattina kannattaa tällaisessa tapauksessa käyttää $\hat{\eta}_p^2$:tä (Cohen, 1973) tai $\hat{\omega}_p^2$:tä (Keren & Lewis, 1979) jotka ovat vertailukelpoisia eri tutkimusasetelmien yli. Käytettäessä näitä estimaatteja tutkimusasetelmassa olevien, tarkastelun ulkopuolella olevien tekijöiden vaikutukset voidaan yksinkertaisesti osittaa pois efektistä:

$$(1.16) \quad \hat{\eta}_p^2 = \frac{SS_{effect}}{SS_{effect} + SS_{s/solut}}$$

$$(1.17) \quad \hat{\omega}_p^2 = \frac{SS_{effect} - df_{effect} MS_{s/solut}}{SS_{effect} + N(n - df_{effect}) MS_{s/solut}}$$

Näiden tunnuslukujen ongelma on kuitenkin se, että täysin satunnaistetussa asetelmassa lohkotekijä pienentää SS_{WG}:tä. Tällöin $\hat{\eta}_p^2$ ja $\hat{\omega}_p^2$ saattavat olla paljon suurempia sellaisissa tutkimuksissa missä on mukana lohkotekijä (Cohen, 1973). Olejnik ja Algina (2003) suosittelevatkin, että tällaisissa asetelmissa käytettäisiin osittaisten $\hat{\eta}_p^2$ - ja $\hat{\omega}_p^2$ -estimaattien sijaan yleistettyjä $\hat{\eta}_p^2$- ja $\hat{\omega}_p^2$ -estimaatteja, jotka ovat vertailukelpoisia yli eri tutkimusasetelmien. Heidän esittämänsä efektikokoparametri on muotoa

$$(1.18) \text{ efekti} = \frac{\sigma^2_{effect}}{\delta \times \sigma^2_{effect} + \sigma^2_{BS}}$$

missä σ^2_{BS} on yksilöiden välinen variaatio, σ^2_{efekti} on tarkasteltavan ANOVA-mallin havaittu efekti ja d = 1, jos mallissa on vain yksi selitettävä muuttuja, muulloin d = 0.

Koska tällä tavoin määriteltyjen efektikoon estimaattien laskukaavat ovat riippuvaisia käytetyistä tutkimusasetelmissa, en esittele niitä tässä yhteydessä vaan kehotan lukijaa tutustumaan Olejnikin ja Alginan (2003) alkuperäisartikkeliin.

EFEKTIKOON SOVELLUKSIA JA ONGELMIA

Efektikoon estimaatti kvantifioi ja standardoi tutkimuksessa havaitun ilmiön amplitudin. Koska efektikoko ei myöskään ilmaise todennäköisyyttä nollahypoteesin paikkansapitävyydelle eikä niiden tulkitsemiseen tarvita arbitraarisia kriittisiä rajoja, efektikokojen raportoimisella voidaan ratkaista johdannossa esitetyt NHST:iin liittyvät ongelmat ainakin tyydyttävästi. Tämän lisäksi efektikoon estimaateilla on myös muita käyttökelpoisia ominaisuuksia, joista eräs hyödyllisimmistä on yhdistettävyys. Useamman saman- tai erilaiseen tulokseen päätyneiden tutkimusten *p*-arvojen yhdistäminen ei lisää ilmiötä koskevaa tietoa kovinkaan paljon. Tällaisesta lähestymistapaa käyttäen voidaankin lähinnä laskea, kuinka moni tutkimus on tuottanut tukea vaihtoehtoiselle- ja kuinka moni nollahypoteesille ja siten tehdä päätelmiä ilmiön toistuvuudesta. Jos tutkijalla on kuitenkin käytettävissään kahdessa tai useammassa riippumattomassa tutkimuksessa havaitut efektikoot, näiden yhdistelmänä voidaan laskea painotettu yhdistetty efektikoko $efekti_m$. Tämän avulla voidaan arvioida kuinka voimakas riippumattoman ja riippuvan muuttujan välinen yhteys on ollut, kun jokaisessa tutkimuksessa havaittu efekti painotetaan tutkimuksen otoskoolla ja tämän jälkeen efektikoot yhdistetään. Lisäksi $efekti_m$:lle voidaan laskea luottamusväli. Menettely on erittäin käyttökelpoinen meta-analyyseissa. Efektikokoja yhdistettäessä on yksinkertaisinta käyttää *r*-estimaatteja. Tällöin jokainen *r* transformoidaan Fisherin *Z*-*r* -muunnoksella

$$(1.19)\quad Z_r = \frac{1}{2}\log_e \frac{1+r}{1-r}$$

lasketaan painotettu *Zr*-keskiarvo ja transformoidaan tämä *r*-*Z* -muunnoksen avulla r_m:ksi. Lopuksi voidaan vielä laskea 95 %:n luottamusväli r_m:lle (Rosenthal & Rubin, 1989). Menettely on erittäin käyttökelpoinen, koska nyt voidaan laskea r_m usealle eri käsittelylle ja luottamusvälejä tarkastelemalla voidaan arvioida poikkeavatko eri käsittelyjen tuottamat efektit toisistaan (Raghunathan, Rosenthal, & Rubin, 1996). Tällaista menettelyä käyttäen tieteellinen psykologinen tieto kumuloituu aidosti, numeerisesti, eikä ainoastaan laadullisesti.

Jos efektikoot on laskettu tutkimuksista, joissa on käytetty erilaisia koeasetelmia, menettely ei kuitenkaan ole yhtä triviaali. Lohkotekijöiden lukumäärä nimittäin vaikuttaa *r*, η- ja ω-tyyppisten efektikoon estimaattien laskemiseen. Vaikka havaitut ryhmäkeskiarvojen erotukset olisivatkin samansuuruiset kahdessa tutkimuksessa, efektit saattavat olla erisuuruiset, koska asetelman yksittäisten solujen varianssi on riippuvainen lohkotekijöiden määrästä (Olejnik & Algina, 2003). Tällöin kannattaakin laskea mieluummin $\hat{\eta}^2_G$- tai $\hat{\omega}^2_G$ -estimaatti jokaiselle tutkimukselle, mutta näiden luottamusvälien muodostaminen ei valitettavasti ole yhtä yksinkertaista kuin r_m:n tapauksessa.

Efektikoon avulla voidaan suorittaa helposti myös voimalaskelmia (ks. Cohen, 1992), joiden avulla voidaan estimoida kuinka suuri otoskoon tulisi olla, jotta tietyn suuruinen efekti olisi tilastollisesti merkitsevä. Odotettu efektin koko voidaan arvioida tarkastelun kohteena olevaa ilmiötä selvittäneiden tutkimusraporttien perusteella yhdistämällä efektikoot siten, kuten edellä esitettiin. Menettely on käytännöllinen, sillä voimalaskelmien avulla voidaan välttyä sekä liian pieniltä että tarpeettoman suurilta otoksilta, ja voimalaskelmien tekemistä voidaankin suositella rutiinitoimenpiteenä mitä tahansa tutkimusta suunniteltaessa.

Kääntäen efektikoon avulla on myös mahdollista laskea niin sanotun pöytälaatikkoefektin (file drawer effect) suuruus (Rosenthal & DiMatteo, 2001). Tällä viitataan siihen, että tutkimuksia, joissa nollahypoteesi jää voimaan, on tyypillisesti hankala saada julkaistuksi. Siten raportoiduista tutkimuksista lasketut r_m:*t* ovat tyypillisesti r_{m_μ}:n yliestimaatteja. Yksinkertainen – ja samalla myös yksittäisten tutkimustulosten efektin suuruutta arvioitaessa käyttökelpoinen menettely – on arvioida, kuinka monta nollatulosta tuottavaa tutkittavaa täytyisi testata, jotta havaittu efektikoko alittaisi ennalta määrätyn rajan (Rosenthal, 1995) tai että *p*-arvo ylittäisi asetetun alfatason (Nummenmaa & Niemi, 2004). Jälkimmäisessä lähestymistavassa määritetään pienin havaittuun dataan yhdistettävä *n* nollatuloksen tuottavia koehenkilöitä jolla 95 %:n luottamusväli r_m:lle käsittää nollan. Tällöin suoritetaan eräänlainen käänteinen voimalaskelma jonka avulla on käytännöllistä arvioida kuinka yleistettävinä tutkimuksessa havaittuja tuloksia voidaan pitää.

Mitä efektikoon estimaattia pitäisi käyttää?

Tutkimusraporttien tulkitsemisen kannalta on ongelmallista, että toistaiseksi ei ole saavutettu yksimielisyyttä siitä, mitä efektikoon estimaattia raportoinnissa tulisi käyttää. Tämä ei kuitenkaan ole ylitsepääsemätön ongelma, sillä *d*-ja *r*-tyyppiset efektikoot voidaan kuitenkin muuntaa jälkeenpäin toisikseen. Jos mahdollista, yleensä kannattaisi kuitenkin käyttää $\hat{\eta}_G^2$ - tai $\hat{\omega}_G^2$ -estimaattia, koska tällaiset estimaatit voidaan laskea useimmissa tutkimusasetelmissa ja siten efektikokojen estimaatit ovat vertailukelpoisia yli erilaisten tutkimusasetelmien. Valitettavasti $\hat{\eta}_G^2$ - ja $\hat{\omega}_G^2$ -estimaattien ongelma kuitenkin on, että heikot efektit saattavat hävitä kun efektikoon estimaatti neliöidään.

Jos tarkasteltava efekti on pieni eikä $\hat{\eta}_G^2$- tai $\hat{\omega}_G^2$-estimaattia haluta käyttää, suosittelen efektikoon estimaatiksi *r*:ää ennemmin kuin *d*-perustaista estimaattia jo siitäkin syystä, että *r* on helpommin tulkittavissa. Siinä missä *d* ilmoittaa efektin suuruuden keskihajonnan yksikkönä, saadaan *r*:n avulla helpommin sovellettavissa olevaa tietoa. Tarkastellaan esimerkkinä Baskinin tutkimusryhmän (2003) meta-analyysia, jossa vertailtiin plaseboterapian ja ”oikean” psykoterapian vaikuttavuuksia. Tutkimusta varten vertailtiin kolmea eri ryhmää psykoterapioita: (i) ”oikeita” psykoterapioita, (ii) ”oikeiden” psykoterapioiden kanssa rakenteellisesti erilaisia plaseboterapioita ja (iii) ”oikeiden” psykoterapioiden kanssa rakenteellisesti samanlaisia plaseboterapioita. Tutkimuksessa havaittiin, että *d* kontrastille i-ii oli 0.47 ja kontrastille i-iii 0.149. Keskihajontaan perustuvien efektikoon estimaattien perusteella ”oikeat” terapiat vaikuttaisivat siis toimivan jossain määrin paremmin kuin kumpikaan plaseboterapioista. Jos *d*-tyyppiset efektikoot muutetaan *r*:ksi niin havaitaan, että vastaavat efektikoon estimaatit ovat 0.23 ja 0.07. Kun ”oikean” psykoterapian vaikuttavuutta verrataan sen kanssa rakenteellisesti samanlaiseen plaseboterapiaan, niin koehenkilön saaman terapian tyypin (oikea / plasebo) havaitaan selittävän vain 7 % hoitotuloksesta. Tulosten soveltajalle *r*-tyyppinen estimaatti on siten intuitiivisesti mielekkäämpi tapa tulkita tutkimuksessa havaittu psykoterapian vaikuttavuus.

YHTEENVETO: EFEKTIKOKOJEN TULKITSEMINEN JA RAPORTOIMINEN

Vaikka tilastotieteilijät ovat jo pitkään kehottaneet käyttäytymistieteiden tutkijoita täydentämään tai korvaamaan NHST:n muilla menettelyillä (Chow, 1988), NHST on edelleen vallitseva tapa hypoteesien testaamiseen psykologiassa. Efektikoon rutiininomainen raportoiminen on alkanut muodostua käytännöksi vasta viime vuosina, ja kaikki lehdet eivät sitä vieläkään edellytä. Tutkimuksen efektikoko kannattaa kuitenkin laskea ja raportoida, vaikka kohdelehden käytäntö ei sitä edellyttäisikään. Efektikokojen avulla saadaan tutkimusaineistosta monipuolisempaa informaatiota kuin NHST:lla. Siinä missä NHST antaa binäärisesti tukea joko nolla- tai vaihtoehtoiselle hypoteesille, efektikoko puolestaan antaa suoraan numeerista informaatiota tutkittavien muuttujien välisestä yhteydestä.

Pelkkä efektikoko ei kuitenkaan riitä tutkimustulosten raportoimiseen. Efektikoko kvantifioi tutkittavan ilmiön amplitudin otoksessa, mutta sen avulla ei voi päätellä, aiheutuiko havaittu efekti sattumalta, ts. kuinka todennäköistä on, että vastaava efekti havaitaan toisessa riippumattomassa aineistossa. Tämän vuoksi efektikoon rinnalle on esitetty myös toisentyyppisiä estimaatteja, joissa tutkimustulosten arvioinnin kannalta keskeinen tieto voitaisiin esittää yhden statistikan avulla. Esimerkiksi Killeenin (2005) r_{rep} estimoi todennäköisyyden sille, että havaitun efektin kanssa samansuuntainen efekti toistuu riippumattomassa tutkimuksessa jossa koehenkilömäärä ja otantavirhe ovat alkuperäisen tutkimuksen kaltaisia.

Efektikoko ei myöskään ole täydellinen estimaatti ilmiöiden välisestä riippuvuudesta. Ensinnäkin, efektikoko (eli riippuvan ja riippumattoman muuttujan välisen yhteyden voimakkuus) ei kaikissa tilanteissa ole ekvivalentti teoreettisten konstruktien välillä vallitsevan assosiaation kanssa (Chow, 1988) eikä sitä pidä myöskään tulkita näin. Lisäksi, vaikka efektikoko onkin informatiivinen tapa esittää tutkimuksen tulokset, efektikoko ei välttämättä ole tutkimustulosten soveltajalle merkityksellistä tietoa (Olejnik & Algina, 2000). Esimerkiksi depressiolääkkeitä kustantavaa tahoa ei kiinnosta tietää, että lääkkeiden vs. plasebon

käyttö selittää 50 % hoidettavien potilaiden depression variaatiosta. Rahoittajaa kiinnostaa, mitä hyötyä depressiolääkkeiden käyttämisestä on ollut näille henkilöille verrattuna jonossa odottamiseen.

Ongelmista huolimatta suosittelen, että efektikoon estimaatteja käytettäisiin systemaattisesti tutkimustuloksia raportoitaessa. Efektikoon estimaatin ohella olisi aina kuitenkin suoritettava joko efektikoon luottamusvälin arviointi (Loftus, 1996; Steiger, 2004), aineiston huolellinen graafinen esittäminen (Loftus, 1996) sekä vahva NHST (Cohen, 1994). Jos efektikoon estimaatit valitaan huolellisesti tarkoitustaan vastaavaksi, tulkitaan asianmukaisesti ja niiden rinnalla esitetään riittävästi täydentävää informaatiota, psykologiassa voidaan muodostaa todellisia teorioita kumuloituvan, kvantitatiivisen tutkimustiedon perusteella.

Artikkeli on saapunut toimitukseen 8.8.2005 ja hyväksytty julkaistavaksi 29.9.2005.

Lähteet

American Psychological Association. (2001). Publication manual of the American Psychological Association (5.th ed.). Washington: American Psychological Association.

Baskin, T. W., Tierney, S. C., Minami, T. & Wampold, B. E. (2003). Establishing specificity in psychotherapy: A meta-analysis of structural equivalence of placebo controls. Journal of Consulting & Clinical Psychology, 71, 973–979.

Bayes, T. (1764). An essay toward solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London, 53, 370–418.

Chow, S. L. (1988). Significance test or effect size? Psychological Bulletin, 103, 105–110.

Cohen, J. (1994). The earth is round (p<.05). American Psychologist, 49, 997–1003.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155–159.

Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor ANOVA designs. Educational and Psychological Measurement, 33, 107–112.

Cohen, J. (1969). Statistical power analysis for the behavioral sciences. New York: Academic Press.

Fisher, R. A. & Bennett, J. H. (1990). Statistical methods, experimental design, and scientific inference. New York: Oxford University Press.

Gill, J. (2002). Bayesian methods: A social and behavioral sciences approach. Boca Raton: Chapman & Hall/CRC.

Glass, G. V., McGaw, B.,& Smith, M. L. (1981). Meta-analysis in social research. Beverly Hills: Sage Publications.

Keren, G. & Lewis, C. (1979). Partial omega squared for ANOVA designs. Educational & Psychological Measurement, 39, 119–128.

Killeen, P. R. (2005). An Alternative to Null-Hypothesis Significance Tests. Psychological Science, 16, 345–353.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56, 746–759.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. Psychological Science, 5, 161–171.

Muthèn, B., & Muthèn, L. K. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. Alcoholism: Clinical & Experimental Research, 24, 882–891.

Nummenmaa, L. (2004). Käyttäytymistieteiden tilastolliset menetelmät. Vammala: Tammi.

Nummenmaa, L., & Niemi, P. (2004). Inducing affective states with success-failure manipulations: A meta-analysis. Emotion, 4, 207–214.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. Psychological Methods, 8, 434–447.

Olejnik, S. & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. Contemporary Educational Psychology, 25, 241–286.

Popper, K. R. (1959). The logic of scientific discovery. New York: Basic Books.

Raghunathan, T. E., Rosenthal, R. & Rubin, D. B. (1996). Comparing correlated but nonoverlapping correlations. Psychological Methods, 1, 178–183.

Rosenthal, R. (1984). Meta-analytic procedures for social research. Beverly Hills: Sage Publications.

Rosenthal, R. (1995). Writing meta-analytic reviews. Psychological Bulletin, 118, 183–192.

Rosenthal, R. & DiMatteo, M. R. (2001). Meta analysis: Recent developments in quantitative methods for literature reviews. Annual Review of Psychology, 52, 59–82.

Rosenthal, R. & Rosnow, R. L. (1991). Essentials of behavioral research: Methods and data analysis. New York: McGraw-Hill.

Rosenthal, R. & Rubin, D. B. (2003). r-sub(equivalent): A simple effect size indicator. Psychological Methods, 8, 492–496.

Rosenthal, R. & Rubin, D. B. (1989). Effect size estimation for one-sample multiple-choice-type data: Design, analysis, and meta-analysis. Psychological Bulletin, 106, 332–337.

Steiger, J. H. (2004). Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. Psychological Methods, 9, 164–182.

Stone, J. V. (2002). Independent component analysis: An introduction. Trends in Cognitive Sciences, 6, 59-64.

Tukey, J. W. (1991). The philosophy of multiple comparisons. Statistical Science, 6, 100–116.