UNIVERZA V LJUBLJANI
FAKULTETA ZA DRUŽBENE VEDE

Urška Reja

Mentor: doc. dr. Gregor Petrič
Somentor: prof. dr. Andrej Blejec

# Načrtovanje učinkovite večnivojske raziskave: simulacijska študija

Magistrsko delo

Ljubljana, 2010

**Načrtovanje učinkovite večnivojske raziskave**: simulacijska študija

V pričujoči magistrski nalogi bom preučevala načrtovanje večnivojskih raziskav, ki se pogosto uporabljajo v družboslovnem raziskovanju, a še preredko glede na pogostost situacij, ko bi bil tovrsten načrt bolj ustrezen in najbolj veljaven. Odnos med načrtom in učinkovitostjo v večnivojskih raziskavah postane manj razviden, pomemben premislek pa predstavlja tudi izbor števila in velikosti skupin, ki naj bi jih vzorčili, da bi zagotovili zadostno učinkovitost raziskave, ki jo določajo preciznost, pristranskost in/ali statistična moč. Raziskovalci so sicer že razvili nekaj programskih orodij za optimizacijsko načrtovanja večnivojskih raziskav, vendar pa gre za orodja, ki se osredotočajo na specifična vprašanja in ne omogočajo celovitejšega pristopa k načrtovanju učinkovitega raziskovalnega načrta. Namen pričujoče magistrske naloge je preseči partikularnost s celovitejšo simulacijsko študijo za načrtovanje učinkovite večnivojske raziskave. Konkretneje bom odgovorila na vprašanje, kako je učinkovitost načrta večnivojske raziskave v smislu predvidene statistične moči in natančnosti ocenjevanja parametrov odvisna od vzorčenja na prvem in drugem nivoju, pri čemer bom preverjala tudi vpliv različnih dejavnikov, kot so različne porazdelitve varianc med nivoji (različni znotrajrazredni koeficienti), neuravnoteženi vzorčni načrti, spreminjanje velikosti vpliva. Poleg vsega naštetega bom upoštevala tudi stroškovni vidik. Oblikovala bom torej integralno simulacijsko študijo, ki ni zgolj ponovitev obstoječih simulacijskih študij, poleg tega pa bom poleg natančnosti ocen preverila tudi statistično moč, ki v konkretni navezi še ni bila eksplicitno objavljena.

***Ključne besede:*** *večnivojska raziskava, moč statističnega testa, natančnost ocenjevanja parametrov, velikost vzorca, Monte Carlo simulacija*

**Designing efficient multilevel studies:** simulation study

In the present work the design of multilevel research is studied that is often used in social sciences but still too rarely regarding the frequency of situations where such a design would be most appropriate and most effective. Important consideration in this aspect would be the selection of the number and size of groups that should be sampled to ensure adequate efficiency of research. In such a way precision, bias and/or statistical power would be easily determined. Researchers have already developed some software tools for optimization of designing multilevel research, but their focus is on specific issues that do not allow a more integrated approach to design an efficient research study. The purpose of this thesis is to overcome particularity by a more comprehensive simulation study to design an effective multilevel research. More specifically, I will answer the question how the design of efficient multilevel study (in the sense of the statistical power and precision of parameter estimates) depends on sampling on the first and on the second level. Furthermore, I will also review the impact of various factors, such as different distribution of the variances between levels (different intraclass coefficients), an unbalanced sampling design and the sizing effect. In addition to all of the above I will also consider the cost aspect. I will therefore form an integral simulation study that is not merely a repetition of existing simulation studies. Besides the accuracy of the estimates I will also examine the statistical power that has not been explicitly published in such a specific relationship.

***Key words:*** *multilevel analysis, power of a statistical test, accuracy in parameter estimation, sample size, Monte Carlo simulation*

# KAZALO

# 1 Uvod

Večnivojski (*angl. multilevel*) raziskovalni načrti se pogosto uporabljajo v družboslovnem raziskovanju, a še preredko glede na pogostost situacij, ko bi bil tovrsten načrt bolj ustrezen in najbolj veljaven. Hierarhični linearni modeli (*angl. hierarchical linear models*), ki združujejo tako slučajne kot fiksne vplive, zagotavljajo uporabno statistično paradigmo v situacijah, kjer je gnezdenje očitna in neposredna posledica večstopenjskega vzorčenja (*angl. multistage sampling*), kot tudi posledica situacij z gnezdenemi viri slučajne variabilnosti (Raudenbush in Bryk 2002; Goldstein 2003; Hox 2002; Snijders in Bosker 1999). Večnivojski statistični modeli se uporabljajo vedno pogosteje, ker na eleganten način obravnavajo odvisnosti med opazovanimi enotami, do katerih pride zaradi večstopenjskega vzorčenja. Do odvisnosti med vzorčenimi opazovanimi enotami prihaja zaradi pripadnosti enot (višjenivojskim) skupinam, ali pa zaradi ponovljenih merjenj, z vrsto merjenih opazovanj znotraj posameznih enot. Za raziskovanje v družboslovju je prvo še posebej pomembno, saj so enote pogosto gnezdene v višje nivojske skupine (npr. učenec-razred-šola, podjetje-panoga, posameznik-družbena skupina), kar pa najpogosteje uporabljene pojasnjevalne metode (npr. regresijska analiza) pogosto zanemarjajo. Podobno lahko z uporabo večnivojskih analiz obravnavamo tudi meta analize in multivariatne analize (Raudenbush 1988; Van den Noortgate in Onghena 2003, 2006).

Da odgovorimo na določeno raziskovalno vprašanje, lahko uporabimo različno učinkovite raziskovalne načrte. Učinkovitost načrta (*angl. design efficiency*) odraža količino informacije, ki jo dobimo iz raziskave v odnosu do stroškov, ki jih raziskava zahteva. Glede na cilj raziskave lahko količino te informacije izmerimo bodisi na podlagi perspektive *ocenjevanja vpliva* in/ali perspektive *testiranja vpliva v hipotezi* (Kelley in Maxwell 2003). Z drugimi besedami to pomeni, da merimo učinkovitost s točnostjo ocene (*angl. accuracy of estimation*) (Kelley in Maxwell 2003) in/ali z močjo statističnega testa (*angl. power of statistical testing*) (J. Cohen 1988), glede na količino sredstev, ki so bila potrebna, da dobimo določeno količino informacije.

Medtem ko je dobro znano kako povečati učinkovitost načrta v enostavnih modelih (npr. Howell 2005), odnos med načrtom in učinkovitostjo postane manj razviden za bolj kompleksne modele in njihove ustrezne analize, kot npr. v analizah večnivojskih podatkov (Snijders 2005). Načrtovanje večnivojskih raziskav predstavlja izredno kompleksen proces, ki zahteva sprejemanje in tehtanje številnih odločitev, ki lahko v končni fazi pomembno vplivajo na kvaliteto/zanesljivost rezultatov. Po mnenju Coolsa in drugih (2008) predstavlja pomemben premislek v načrtovanju večnivojske raziskave izbor števila in velikosti skupin, ki naj bi jih vzorčili, da bi zagotovili zadostno učinkovitost raziskave. Če ponazorim na primeru: ko raziskovalec načrtuje večstopenjsko vzorčenje za analizo vpliva šol na dosežke dijakov, ali testira hipotezo, da so državljani siromašnih soseščin bolj pogosto žrtve kriminala kot drugi državljani, se mora odločati o velikostih skupin na različnih nivojih. Za dvo-nivojski načrt v prvem primeru si lahko postavimo vprašanje: ali naj raziskovalec preučuje veliko šol z manj dijaki, ali malo šol z veliko dijaki, pri čemer seveda predpostavljamo, da obstajajo proračunske omejitve za izvedbo raziskave.

## 1.1 Obstoječi pristopi k raziskovanju učinkovitosti načrta v večnivojskih raziskavah

Obstajajo tako analitični kot simulacijski pristopi k raziskovanju učinkovitosti načrta v večnivojskih raziskavah. Analitični pristopi temeljijo na rabi matematičnih formul, iz katerih je moč izpeljati standardne napake in moč testa za ocenjevanje in testiranje specifičnih koeficientov določenega večnivojskega modela, kar lahko predstavlja hitro sredstvo za obravnavanje cele vrste alternativnih raziskovalnih načrtov. Simulacijski pristopi, na drugi strani, pa predstavljajo uporabo simulacij za specifičen raziskovalni načrt. Oba pristopa imata svoje omejitve. Analitične raziskave tako omogočajo le obravnavo preprostih modelov in lahko obravnavajo/manipulirajo le nekaj predpostavk, rezultate simulacijskih raziskav pa je težko posploševati, saj so podatki generirani za specifičen model in za specifične vrednosti parametrov (Cools in drugi 2008).

Le nekaj raziskav je dejansko privedlo do računalniških aplikacij, ki omogočajo avtomatično izpeljavo analitičnih izračunov za nekatere parametre večnivojske

raziskave (npr. PINt - Power in Two-level designs (Snijders in Bosker 1993), OD - Optimal Design (Raudenbush in Liu 2001)). Za nekatere raziskovalne situacije pa je izpeljava zahtevanih formul v zaprti obliki nemogoča, zato so mnogi raziskovalci le-to zamenjali s simulacijskimi postopki, ki pa so generirani za specifičen model in za specifične vrednosti parametrov in jih tako ne moremo posploševati. V ta namen sta bili izdelani statistični orodji (ML-DEs (Cools in drugi 2008), MLPowSim (Browne in Golalizadeh 2009)), ki se poslužujeta simulacij in sta precej bolj fleksibilni kot zgoraj omenjeni programski orodji.

## 1.2 Cilj naloge

Raziskovalec z znanjem statistike in osnovami programiranja si lahko za svoj raziskovalni problem postavi lastno simulacijsko študijo za načrtovanje učinkovite večnivojske raziskave, konkretneje za določanje optimalnega števila in velikosti skupin, ki naj bi jih vzorčili, da bi zagotovili zadostno učinkovitost raziskave. Obstoječa programska orodja so namenjena raziskovalcem z malo statističnega znanja in brez osnov programiranja, saj nekaterih stvari ne omogočajo oz. so osredotočena na specifična vprašanja.

Namen pričujoče magistrske naloge je tako preseči omejitve, ki izvirajo iz posameznih pristopov s celovitejšo simulacijsko študijo za načrtovanje učinkovite večnivojske raziskave. S pomočjo simulacijske študije bom definirala optimalno število skupin in velikosti vzorcev z vidika statistične moči in natančnosti ocenjevanja parametrov. Pod določenimi pogoji bom preverjala tako natančnost regresijskih koeficientov in njihovih standardnih napak kot tudi natančnost variančnih komponent in njihovih standardnih napak. Na ta način bom prišla do odgovora, kako je učinkovitost načrta večnivojske raziskave v smislu predvidene statistične moči in natančnosti ocenjevanja parametrov odvisna od vzorčenja na prvem in drugem nivoju, pri čemer bom preverjala tudi vpliv različnih dejavnikov, kot so različne porazdelitve varianc med nivoji (različni znotrajrazredni koeficienti), neuravnoteženi vzorci, občutljivost natančnosti ter moči glede na velikost vpliva. Poleg vsega naštetega pa bom v analizo vključila tudi stroškovni vidik.

Na ta način bom oblikovala simulacijsko študijo, ki ni zgolj ponovitev obstoječih simulacijskih študij, saj gre v mojem primeru za celovitejšo raziskavo, ki poleg kontroliranja velikosti skupin in vzorcev, v pojasnjevanje učinkovitosti načrta večnivojske raziskave vključi tudi številne druge, že omenjene, parametre in dejavnike. Poleg tega pa bom poleg natančnosti ocen preverila tudi statistično moč, ki (kolikor mi je znano) v konkretni navezi še ni bila eksplicitno objavljena.[1]

[1] Seveda pa so zaključki glede natančnosti ocenjevanja in moči statističnega testiranja povezani; vsaka razlika med njimi predstavlja funkcijo velikosti učinka (Kelley in Maxwell 2003). Natančnost odraža količino informacije, ki jo dobimo iz vrednosti populacijskega parametra, medtem ko moč odraža količino informacije, ki jo dobimo z zavračanjem ničelne hipoteze. Večja natančnost pomeni večjo moč, ampak optimalna natančnost ne pomeni nujno optimalne moči, ker je lahko obravnavani vpliv izredno majhen.

## 1.3 Struktura magistrske naloge

V drugem poglavju bom obravnavala učinkovitost raziskovalnega načrta, ki jo lahko definiramo kot količino informacije, ki jo dobimo iz raziskave, glede na stroške, ki jih raziskava zahteva. Glede na cilj raziskave lahko količino te informacije izmerimo bodisi na podlagi perspektive testiranja vpliva v hipotezi in/ali perspektive ocenjevanja vpliva.

V naslednjem poglavju bom na kratko predstavila večnivojsko analizo in hierarhične linearne modele. Konkretneje bom predstavila enostaven model enostranske analize variance ANOVA s slučajnimi vplivi ter model s slučajnimi presečišči in nagibi, teorijo o ocenjevanju parametrov ter testiranju hipotez.

Sledi poglavje o načrtovanju večnivojske raziskave, ki jo lahko obravnavamo z vidika analitične metodologije ali pa z vidika simulacijskih raziskav. V tem poglavju bom predstavila prednosti ter slabosti obeh načinov.

V šestem poglavju bom obravanavala programe, ki so na voljo za določanje učinkovitih načrtov večnivojskih raziskav. Nekateri se poslužujejo analitične, drugi simulacijske metodologije.

Sedmo poglavje obravnava učinkovito načrtovanje velikosti vzorca na osnovi simulacij, v osmem poglavju so predstavljeni rezultati simulacijske študije, v zaključnem poglavju pa so predstavljene ključne ugotovitve simulacijske raziskave.

# 2 Učinkovitost raziskovalnega načrta

Učinkovitost raziskovalnega načrta (*angl. design efficiency*) lahko definiramo kot količino informacije, ki jo dobimo iz raziskave v odnosu do stroškov, ki jih raziskava zahteva. Glede na cilj raziskave lahko količino te informacije izmerimo bodisi na podlagi perspektive *testiranja vpliva v hipotezi* ali perspektive *ocenjevanja vpliva* (Kelley in Maxwell 2003).

Raziskovalci lahko pri načrtovanju raziskave izračunajo oceno velikosti vzorca iz perspektive analize moči statističnih testov z namenom, *da bi dosegli neko smiselno verjetnost za dosego statistično značilnih ocen parametrov*. V splošnem se družboslovna znanost vedno bolj zaveda problemov, povezanih z raziskavami s premajhno statistično močjo in njim ustreznimi napakami Tipa II, ki lahko vodijo v napačne rezultate v danem območju raziskovanja (J. Cohen 1994; Muller in Benignus 1992; Sedlmeier in Gigerenzer 1989). V tem primeru g*re za perspektivo testiranja vpliva v hipotezi.*

Izvedbo načrtovanja velikosti vzorca samo za dosego statistično značilne ocene parametra lahko pogosto izboljšamo z načrtovanjem velikosti vzorcev, ki vodijo do statistično značilnih in natančnih ocen parametrov (*angl. accurate parameter estimates*) (Kelley in Maxwell 2003). *V tem primeru gre za perspektivo ocenjevanja vpliva.*

Seveda pa v večjih raziskavah želimo testirati veliko hipotez in več parametrov, ki jih moramo obravnavati posebej. Izračunavanje velikosti vzorca tako obravnavamo kot groba vodila, saj vedno obstaja negotovost glede pravih ocen, pogosto pa nas omejujejo tudi praktične omejitve, ki jih moramo upoštevati.

V nadaljevanju bom predstavila obe perspektivi, povezani z načrtovanjem večnivojske raziskave, perspektivo testiranja vpliva v hipotezi in perspektivo ocenjevanja vpliva.

## 2.1 Perspektiva testiranja v hipotezi: moč statističnega testa

V najbolj osnovni obliki ocenjujemo le velikost vzorca, ki jo zahteva testiranje hipoteze (Browne in drugi 2009). Z vidika statističnega testiranja informacijo izmerimo s statistično močjo. Splošno predstavitev analize statistične moči lahko najdemo v delu Cohena (1988), ali za hiter uvod, delo istega avtorja (1992).

### 2.1.1 Testiranje hipoteze

Ko se raziskovalec odloči za raziskovanje določenega področja, ima običajno pred seboj neko raziskovalno vprašanje. Tako ga lahko npr. primarno zanima kateri dejavniki vplivajo na dosežke dijakov ob koncu šolanja. To splošno raziskovalno vprašanje se lahko razdeli v nekaj specifičnih hipotez, npr. *'fantje v povprečju dosežejo slabše rezultate, če obravnavamo skupni dosežek v starosti 16 let'*, ali podobno *'dekleta dosegajo boljše rezultate kot fantje'*.

Za testiranje prve hipoteze bi moral raziskovalec ugotoviti kakšen je dosežek za slučajni vzorec fantov v starosti 16 let, prav tako pa bi moral poznati tudi splošno povprečno vrednost dosežka za vse dijake. Da bi dobili razliko med tema dvema vrednostma, bi morali vzorčno povprečje za fante primerjati s splošnim povprečjem, pri čemer bi uporabili velikost vzorca in variabilnost v dobljenem dosežku fantov. Na ta način bi ugotovili, ali je ta razlika več kot bi lahko pričakovali po naključju. V primeru druge hipoteze bi lahko za slučajni vzorec fantov in deklet, starih 16 let, izmerili skupni dosežek in primerjali vzorčni povprečji po spolu, nato pa bi z uporabo velikosti vzorcev in variabilnosti določili, ali je razlika večja kot bi pričakovali po naključju.

Našo začetno oz. alternativno hipotezo ($H_1$) (*'fantje v povprečju dosežejo slabše rezultate od skupnega povprečja')* primerjamo z ničelno hipotezo ($H_0$)[2] (*'fantje se v dosežkih ne razlikujejo od skupnega povprečja')*. Denimo, da smo transformirali podatke tako, da znaša skupno povprečje 0.

[2]Tako imenovano zato, ker izniči raziskovalno vprašanje, ki ga želimo dokazati.

Želimo torej testirati hipotezi: $H_0$: $\mu_F = 0$ *proti* $H_1$: $\mu_F < 0$, kjer $\mu_F$ predstavlja skupno povprečje za celotno populacijo fantov (populacijsko povprečje). Potrebujemo neko pravilo oz. kriterij za odločanje med tema dvema hipotezama. V tem primeru bi bilo smiselno obravnavati vrednost vzorčnega povprečja $\bar{x}$ in nato zavrniti ničelno hipotezo pri $\bar{x} \leq c$, kjer je *c* neka izbrana konstanta. Če je $\bar{x} > c$, $H_0$ ne moremo zavrniti, saj ne moremo z dovoljšnjo gotovostjo trditi, da fantje v resnici dosegajo slabše rezultate od povprečja. Izračunati je potrebno mejno vrednost *c*, kjer se bo naša odločitev spremenila.

### 2.1.2 Moč statističnega testa

Moč statističnega testa se nanaša na verjetnost zavrnitve ničelne hipoteze, ko ta dejansko ne drži (Cohen, 1992). Želimo potrditi raziskovalno hipotezo ($H_1$), ki trdi, da določen vpliv obstaja, in tako testiramo ničelno hipotezo o odsotnosti tega vpliva ($H_0$) z uporabo vzorca iz relevantne populacije. Statistična značilnost α predstavlja tveganje za napačno zavrnitev $H_0$. Ta napaka je znana kot napaka Tipa I. *Po drugi strani* β predstavlja verjetnost, da ne zavrnemo $H_0$ v primeru, da vpliv v resnici obstaja v populaciji. Ta napaka je znana kot napaka Tipa II (glej Tabelo 2.1). Statistična moč testa predstavlja verjetnost zavrnitve $H_0$ glede na dano velikost vpliva v populaciji, stopnjo značilnosti α ter velikost vzorca v raziskovalnem načrtu. Moč je tako opredeljena kot 1- β.

**Tabela 2.1: Vrste napak**

| | DEJANSKO STANJE | | |
|---|---|---|---|
| | | Pravilna $H_0$ | Napačna $H_0$ |
| UGOTOVITVE RAZISKAVE | Pravilna $H_0$ | O.K. | **napaka Tipa II** |
| | Napačna $H_0$ | **napaka Tipa I** | O.K. |

Imamo torej količino *c*, ki jo lahko prilagodimo za določen vzorec, pri čemer pa ne moremo kontrolirati vrednosti obeh, α in β. V splošnem izberemo vrednost *c*, ki nam omogoča, da dobimo določeno vrednost za α. Če lahko predvidevamo določeno obliko porazdelitve za vzorčno povprečje (oz. funkcijo le-tega), potem lahko za ugotavljanje verjetnosti zavrnitve $H_0$ za različne vrednosti *c*, uporabimo lastnosti te porazdelitve. V našem primeru bomo predvidevali, da dosežek za vsakega posameznega dijaka (fanta) ($x_i$) prihaja iz normalne porazdelitve s povprečjem $\mu_F$ in neznano varianco $\sigma^2_F$. Če bi poznali varianco, bi lahko predpostavljali, da tudi

vzorčno povprečje prihaja iz normalne porazdelitve s povprečjem $\mu_F$ in varianco $\sigma^2_F / n$, pri čemer n predstavlja velikost našega vzorca.

Tako lahko trdimo, da $\frac{\bar{x} - \mu_F}{\sigma_F / \sqrt{n}}$ sledi standardizirani normalni porazdelitvi. Če želimo izračunati verjetnost $P(\bar{x} \le c) = \alpha$, potem $P(\bar{x} \le c) = P\left[\frac{\bar{x} - \mu_F}{\sigma_F / \sqrt{n}} \le \frac{c - \mu_F}{\sigma_F / \sqrt{n}}\right] = \alpha$ privede do $\frac{c - \mu_F}{\sigma_F / \sqrt{n}} = Z_\alpha$, kjer je $Z_\alpha$ α-ti kvantil normalne porazdelitve. Prestavitev členov nam da enačbo $c = \mu_F + Z_\alpha \sigma_F / \sqrt{n}$.

Običajno variance $\sigma^2_F$ ne poznamo in jo zamenjamo z vzorčno varianco $s^2_F$ ampak ker je to ocena za $\sigma^2_F$, moramo v obzir vzeti tudi njeno porazdelitev. Tako namesto normalne porazdelitve uporabimo $t_{n-1}$ porazdelitev, kar privede do temu ustrezno spremenjene formule za mejno vrednost (*angl. treshold*) $c = \mu_F + t_{n-1,\alpha} s_F / \sqrt{n}$. Ko velikost vzorca narašča, se t-porazdelitev približuje normalni porazdelitvi, zato pogosto kot približek kvantilov t-porazdelitve uporabimo kar kvantile normalne porazdelitve.

Moč testa za določeno vrednost $\mu_F$ lahko ovrednotimo. Če npr. verjamemo, da je prava vrednost $\mu_F = -1$, potem lahko ocenimo moč testa glede na to dano vrednost. Zanima nas, kako pogosto lahko zavrnemo ničelno hipotezo, če bi bila določena alternativa $\mu_F = -1$ v resnici prava. Imamo $Moč = P(\bar{x} \le c \,|\, \mu_F = -1)$, pri čemer *c* izračunamo pod ničelno hipotezo, to je:

$$Moč = t^{-1}_{n-1}\left(\frac{c+1}{s_F / \sqrt{n}}\right) = t^{-1}_{n-1}\left(\frac{(t_{n-1,\alpha/2} s_F / \sqrt{n}) + 1}{s_F / \sqrt{n}}\right).$$

Npr. če je velikost vzorca, n=100, vzorčna varianca $s_F = 1$ in α=0.05 (2-stranski), imamo približno $t_{99,\ 0.05/2}$ = -1.98 in $Moč = t^{-1}_{99}((-0.198+1)/0.1) = t^{-1}_{99}(8.02) \approx 1$ (ogromna). V tem primeru je 100 fantov več kot dovolj za veliko moč.

Če verjamemo, da je prava vrednost $\mu_F = -0.10$, potem izračunamo

$Moč = t_{99}^{-1}((-0.198 + 0.10)/0.1) = t_{99}^{-1}(-0.98) = 0.165$.

V tem primeru je moč nizka in bi potrebovali večji vzorec, da bi zagotovili zadostno moč. Če bi želeli izračunati velikost vzorca, ki bi nam dala moč 0.8, bi morali rešiti enačbo za n; to je težje v primeru t-porazdelitve kot v primeru normalne porazdelitve, saj se porazdelitvena funkcija t spreminja z n. Upoštevamo lahko tudi, da se z rastočim n, t-porazdelitev vedno bolj približuje normalni porazdelitvi.

Če torej v našem primeru predpostavljamo normalno porazdelitev, imamo nekoliko bolj enostavno formulacijo:

$$Moč = \phi\left(\frac{c + 0.1}{s_F/\sqrt{n}}\right) = \phi\left(\frac{(Z_{\alpha/2} s_F/\sqrt{n}) + 0.1}{s_F/\sqrt{n}}\right),$$

pri čemer $\phi = Z^{-1}$ predstavlja inverz standardne normalne kumulativne funkcije (CDF).

V primeru, kjer je $s_F = 1$ in $Z_{\alpha/2=} -1.96$, imamo

$Moč = \phi\left[\frac{(-1.96/\sqrt{n}) + 0.1}{1/\sqrt{n}}\right]$, kar pomeni, da za Moč vsaj 0.8, dobimo:

$$\phi\left[\frac{(-1.96/\sqrt{n}) + 0.1}{1/\sqrt{n}}\right] \geq 0.8 \rightarrow \frac{(-1.96/\sqrt{n}) + 0.1}{1/\sqrt{n}} \geq 0.842.$$

Reševanje enačbe po n nam da $n \geq (10 * (0.842 + 1.96))^2 = 785.1$. Potrebovali bi vzorec vsaj 786. Tu predstavlja vrednost 0.842 vrednost repa normalne porazdelitve (*angl. tail of the normal distribution*), povezane v močjo 0.8 (nad katero leži 20% porazdelitve).

### 2.1.3 Zakaj je Moč pomembna?

Ko želimo odgovoriti na določeno raziskovalno vprašanje upamo, da je ničelna hipoteza napačna in da jo bomo na osnovi naših podatkov lahko zavrnili. Če naša domnevna populacijska vrednost predpostavlja test hipoteze z nizko močjo, kljub pravilnosti naše alternativne hipoteze pogosto ne bo mogoče zavrniti ničelne hipoteze. Z drugimi besedami; lahko zapravimo kar nekaj denarja z zbiranjem podatkov v poskusu, da ovržemo ničelno hipotezo, pa nam to ne uspe.

### 2.1.4 Elementi Moči

Analiza statistične moči uporablja odnose med štirimi spremenljivkami, ki so vpletene v statistično sklepanje: velikost vzorca (n), izbrana raven tveganja (α), populacijska velikost vpliva in statistična moč (1- β). Za katerikoli statistični model je vsak element funkcija ostalih treh. Tako lahko npr. za vsak dan statistični test določimo moč za dan α, velikost vpliva ter n. Za načrtovanje raziskave je tako najbolje določiti tolikšno velikost vzorca, da dobimo zaželeno moč glede na α in velikost vpliva. V nadaljevanju so ti elementi podrobneje opisani.

***Moč*** Za splošno uporabo se priporoča specifikacija moči .80 (β=.20) oz. .90 (β=.10). Tako bomo v 80% oz. 90% primerih zavrnili ničelno hipotezo (kar je odvisno od natančnosti naših pravih ocen). Bistveno manjša vrednost od .80 bi privedla do velikega tveganja za Napako tipa II, bistveno večja vrednost, pa do velikosti vzorca, ki bi zelo verjetno prekoračila sredstva, ki jih ima na voljo raziskovalec za izvedbo raziskave.

***Izbrana raven tveganja (α)*** Gre za tveganje, da zmotno zavrnemo ničelno hipotezo ($H_0$) in tako napravimo napako Tipa I, α. Ali bolj opisno; gre za verjetnost, da mislimo, da smo nekaj našli, čeprav to v resnici ne obstaja. Če ni drugače rečeno (in redko je), je določena na .05, seveda pa lahko izberemo tudi druge vrednosti.

***Velikost vzorca (n)*** Pri načrtovanju raziskave mora raziskovalec poznati velikost vzorca (n) in velikosti vpliva v hipotezi. Optimalna velikost vzorca (n) se povečuje s povečanjem želene moči, zmanjšanjem velikosti vpliva in zmanjšanjem α. Za statistični test, ki vključuje dve ali več skupin, tako definiran n predstavlja potrebno velikost vzorca za vsako skupino.

***Populacijska velikost vpliva*** Najtežji del analize statistične moči za raziskovalce je določanje populacijske velikosti vpliva (*angl. effect size*). V Neyman-Pearsonovi metodi statističnega sklepanja je poleg specifikacije $H_0$ le-tej postavljena nasproti alternativna hipoteza ($H_1$). Velikost vpliva si lahko predstavljamo kot raziskovalčevo idejo o 'stopnji, za katero verjame, da je ničelna

hipoteza napačna' (Cohen, 1992: 156). Z drugimi besedami; gre za razliko med $H_0$ in $H_1$. Če predpostavljamo, da je prava vrednost $\mu_F = -1$, ničelna hipoteza pa ustreza $\mu_F = 0$, dobimo velikost vpliva 1[3]. V pristopu, osnovanem na simulaciji, pogosto ta pojem izpustimo in uporabimo pojme kot so ocena parametrov ali ocena regresijskih koeficientov.

Obstaja kar nekaj metod za določanje velikosti vpliva:

- *Sklepanje glede na obstoječe poznavanje področja.* Kraemer in Thiemann (1987) sta podala naslednji primer. Denimo, da predpostavljamo, da bodo imeli štiridesetletni moški, ki spijejo več kot 3 skodelice kave na dan, višji CMI (angl. Cornel Medical Index) kot moški, ki ne pijejo kave. Indeks rangira od 0 do 195, prejšnje raziskave pa so pokazale, da vsake 10 let pri posamezniku indeks naraste za 3.5 točke. Raziskovalci se odločijo, da bo povečanje indeksa zaradi pitja kave, ki bo enako desetletnemu povečanju indeksa, dovolj veliko za 'povzročitev skrbi' in tako izračunajo velikost vpliva na tej predpostavki.
- *Osnovanje na prejšnjih raziskavah.* Raziskovalec pregleda kaj so ugotovile raziskave iz podobnega raziskovalnega področja.
- *Uporaba dogovorov.* J. Cohen (1988, 1992) ter drugi (npr. Murphy in Myors 2003) predlagajo vnaprej dogovorjene indekse za veliko vrsto statističnih testov.

### *Uporaba dogovorov pri določanju populacijske velikosti vpliva*

Vsak statistični test ima svoj indeks velikosti vpliva. Vsi ti indeksi imajo svobodno lestvico in so zvezni, rangirajo pa od nič naprej in za vse je $H_0$ pri velikosti vpliva 0. Da bi prišli do pomena vseh danih indeksov pa moramo imeti neko idejo o njihovih merskih lestvicah, pri čemer Cohen:1992 predlaga lestvico majhne, srednje ter velike vrednosti velikosti vpliva (več o tem v Cohen 1992:156-157).

Če npr. v neki raziskavi raziskovalec verjame, da je populacijski korelacijski koeficient r srednje velikosti (r=.30 iz tabele, Cohen 1992:156) in da bo izveden t-test z dvostranskim α=.05, bo moč statističnega testa .80 pri velikosti vzorca 85 (Tabela 2 v

[3] Običajna praksa je predvidevati pozitivno velikost vpliva.

Cohen 1992:158). Če se pri velikosti vzorca 85 izkaže, da t-test ni statistično značilen, pomeni, da je r manjši od .30, ali pa je bil raziskovalec žrtev 20% tveganja za Napako tipa II (β=.20).

Za mnoge vrste načrtov lahko raziskovalec izbere velikost vzorca, ki je potrebna za dosego določenega nivoja moči na podlagi dela Cohena. Za večnivojske načrte je načrtovanje raziskave precej bolj kompleksno, saj obstajata (vsaj) dve vrsti velikosti vzorcev: velikosti vzorca na mikro nivoju (n) ter velikosti vzorca na makro nivoju (J) z JXn skupno velikostjo mikro enot.

### 2.1.5 Izvedba analize statistične moči

Obstajajo tri vrste analize statistične moči: *a priori*, *post hoc* ter kompromisna. Slednja je bolj kompleksna, nekoliko kontroverzna in redko uporabljena (Murphy in Myors 2003), zato o njej v pričujoči magistrski nalogi ne bomo pisali.

***A priori analiza statistične moči*** V idealnem svetu je analiza statistične moči izvedena znotraj načrtovalne faze raziskave. S takšnim pristopom zmanjšamo verjetnost, da namenimo veliko časa in sredstev za raziskavo, v kateri imamo le majhno verjetnost, da ugotovimo statistično značilen vpliv. Po drugi strani pa zagotavlja tudi, da ne tratimo časa in sredstev s testiranjem večjega števila subjektov kot je potrebno za ugotavljanje nekega vpliva.

***Post hoc analiza statistične moči*** Medtem ko je *A priori* analiza statistične moči izvedena pred izvedbo raziskave, pa je *post hoc* izvedena po izvedbi in nam pomaga razložiti rezultate, če nismo ugotovili statistično značilnih vplivov.

### 2.1.6 Kako izvajamo izračunavanja moči / velikosti vzorcev bolj splošno?

Predvidevamo, da je velikost vpliva izražena z nekim parametrom $\gamma$, ki ga lahko ocenimo z določeno standardno napako, ki jo označimo s SE($\hat{\gamma}$). Pri tem imamo v mislih, da je velikost standardne napake monotona padajoča funkcija velikosti vzorca; večja je velikost vzorca, manjša je standardna napaka. V večini enonivojskih načrtov je standardna napaka ocene inverzno proporcionalna (ali približno tako) kvadratnemu korenu velikosti vzorca.

Odnos med velikostjo vpliva, stopnjo značilnosti in velikostjo vzorca lahko predstavimo v eni formuli (Snijders in Bosker 1999). Ta formula predstavlja aproksimacijo, ki je veljavna za praktično uporabo, ko želimo uporabiti enostranski t-test za $\gamma$ z dovolj velikim številom prostostnih stopenj (npr. d.f.≥10). Testno statistiko za t-test lahko izrazimo z razmerjem $t = \hat{\gamma} / SE(\hat{\gamma})$. Formula je:

$$\frac{velikost \quad vpliva}{s\tan dardna \quad napaka} \approx (z_{1-\alpha} + z_{1-\beta}) = (z_{1-\alpha} - z_{\beta}) \qquad \textbf{(2.1)},$$

kjer so $z_{1-\alpha}$, $z_{1-\beta}$ in $z_{\beta}$ z vrednosti (vrednosti iz standardne normalne porazdelitve) povezane z ustreznimi vrednostmi kumulativne porazdelitve.

Če npr. izberemo α=.05 in 1-β=.80 (tako da je β=.20) in pričakujemo velikost vpliva .50, lahko izpeljemo, da iščemo minimalno velikost vzorca, ki ustreza:

$$s\tan dardna \quad napaka \leq \frac{.50}{1.64 + .84} = .20 .$$

Formula vsebuje štiri 'neznanke'. To pomeni, da če so dane 3, lahko izračunamo četrto. V praksi je stopnja značilnosti α dana vnaprej, velikost vpliva hipotetično pretehtana (oz. o njej ugibamo), ali pa je znana standardna napaka in je izračunana moč testa 1-β, oz. je znana statistična moč in je izračunana standardna napaka. Glede na dano standardno napako, potem poskušamo izračunati zahtevano velikost vzorca.

## 2.2 Perspektiva ocenjevanja vpliva: AIPE (*angl. Accuracy in Parameter Estimation*)

Natančnost (*angl. accuracy*) bomo obravnavali na dva načina. V najbolj osnovnem smislu natančnost definirata pristranskost ter preciznost, združena v MSE (srednja kvadratna napaka) oz. RMSE (kvadratni koren srednje kvadratne napake (*angl. root mean squared error*), v širšem smislu pa je obravnavana v smislu preciznosti, ki definira širino intervala zaupanja okoli relevantnega parametra. Osnovna ideja načrtovanja velikosti vzorca za natančnost je osnovana na nadzorovanju širine intervala zaupanja našega zanimanja.

### 2.2.1 Natančnost in (R)MSE

Natančnost je v najbolj osnovnem smislu definirana kot ujemanje med vrednostjo, ki jo dobimo iz vzorca in dejansko, toda v glavnem neznano populacijsko vrednostjo. Povprečna kvadratna napaka ali MSE cenilca (*angl. estimator*) je eden od mnogih načinov kako ovrednotiti količino, s katero se cenilec loči od prave vrednosti populacijskega parametra. Za nepristranski cenilec, MSE predstavlja varianco. V analogiji s standardnim odklonom, pa kvadratni koren MSE privede do RMSE, ki ima enake enote kot količina, ki jo ocenjujemo. Če je cenilec nepristranski, je RMSE kvadratni koren variance, znan kot standardna napaka (Casella in Lehman 1999).

Formalna definicija natančnosti je dana s kvadratnim korenom srednje kvadratne napake in jo lahko izrazimo z naslednjo formulacijo ($\hat{\Theta}$ naj bo ocena populacijskega parametra $\Theta$):

$$RMSE = \sqrt{E\left[\left(\hat{\Theta} - \Theta\right)^2\right]} = \sqrt{E\left[\left(\hat{\Theta} - E\left[\hat{\Theta}\right]\right)^2\right] + \left(E\left[\hat{\Theta} - \Theta\right]\right)^2} \qquad \textbf{(2.2)},$$

kjer prva komponenta pod drugim korenom predstavlja **preciznost** (*angl. precision*), druga pa **pristranskost** (*angl. bias*). Ko je pričakovana vrednost parametra enaka vrednosti parametra, ki ga predstavlja (gre za nepristransko oceno), sta natančnost ter preciznost ekvivalentna koncepta.

***Pristranskost*** Pristranskost se nanaša na kar nekaj statističnih tematik, ki jih lahko klasificiramo kot pristranskost merjenja, pristranskost vzorčenja in pristranskost ocenjevanja. V situacijah merjenja pristranskost predstavlja 'razliko med populacijskim povprečjem rezultatov merjenja ali rezultatov testa in sprejeto referenco prave vrednosti' (Bainbridge 1985). Tako pristranskost vodi do pod- ali pre-cenjenosti prave vrednosti. Do *pristranskosti zaradi merjenja* pride v glavnem zaradi nepravilnih merskih naprav ali postopkov. V tem smislu pristranskost z večjim vzorcem ne izgine, saj so vse vrednosti merjenja sistematično oddaljene od prave vrednosti v enaki meri (Kotz in Johnson 1982, 1988, Debanne 2000). Do *pristranskosti zaradi vzorčenja* pride zaradi nereprezentativnega vzorčenja ciljne populacije. Tudi ta vrsta pristranskosti ne izgine s povečanim naporom vzorčenja. *Pristranskost zaradi ocenjevanja* imenujemo tudi sistematična napaka in se nanaša na metodo ocenjevanja, za katero se povprečje ponovljenih ocen razlikuje od prave vrednosti (West 1999). Do te pristranskosti pride zaradi samega cenilca, ki je pristranski. V tem primeru naj bi se pristranskost z večanjem vzorca(ev) zmanjševala, saj je to ena od zaželenih lastnosti cenilca. Naj bo $\hat{\Theta}$ ocena populacijskega parametra $\Theta$; potem je pristranskost dana s $\hat{\Theta} - \Theta$.

***Preciznost*** Preciznost se nanaša na odsotnost slučajne napake. Za razliko od pristranskosti je velikost preciznosti odvisna le od ocenjenih (oz. opazovanih) vrednosti in je popolnoma neodvisna od resnične vrednosti. Preciznost je tako mera *'statistične variance postopka ocenjevanja'* (West 1999) ali v situacijah vzorčenja *'razširjenost podatkov…ki jih lahko pripišemo statistični variabilnosti, ki je prisotna v vzorcu'* (Debanne 2000). V situacijah merjenja preciznost izhaja iz variance, ki jo proizvaja merska naprava ali postopek. Preciznost ocene tako predstavlja varianca koeficientov, formalno jo določimo z: $E\left[\left(\hat{\Theta} - E\left[\hat{\Theta}\right]\right)^2\right]$.

## 2.2.2 Natančnost in INTERVALI ZAUPANJA

Začnimo s preprostim primerom. Obravnavajmo primer raziskovalca, ki načrtuje raziskavo z dvema skupinama, pri čemer je cilj primerjati povprečni vrednosti eksperimentalne in kontrolne skupine. Zaradi večje enostavnosti predvidevajmo, da so udeleženci slučajno izbrani v skupino. Nadalje predvidevajmo, da sta normalnost

in homogenost variance verjetni predpostavki, tako da raziskovalec lahko načrtuje analizo teh podatkov s t-testom za neodvisne skupine z dvo-stransko stopnjo značilnosti α=0.05.

Denimo, da želi raziskovalec imeti moč testa 1-β=0.80. Na začetku težavo povzroča določanje velikosti vpliva. Denimo, da se raziskovalec odloči, da bo sledil vodilom Cohena (1988) in na tej osnovi določi srednjo velikost vpliva (to je Cohenov populacijski d=0.50). Raziskovalec ugotovi, da bo za moč 0.80 potreboval 64 udeležencev na skupino ali skupno velikost vzorca 128, pri predpostavki, da ne bo nikakršnega upada enot. Denimo, da raziskovalec izvede raziskavo in znaša standardizirana razlika vzorčnih povprečij med skupinama točno 0.50 in je tako točno srednja velikost glede na Cohenova (1988) pravila. Ustrezna t-vrednost je 2.83, ki je statistično značilna na stopnji 0.05. Na podlagi rezultatov raziskovalec zaključi, da obstaja razlika med skupinama in da razlika ustreza srednji velikosti. Ampak izračunana velikost vpliva je samo ocena in kot taka je odvisna od variabilnosti. Pomembni viri (Am. Psychol. Assoc. 2001; Wilkinson in drugi 1999, American Educational Research Association 2006) zagovarjajo uporabo intervalov zaupanja pri poročanju rezultatov. Denimo, da naš raziskovalec sledi tem napotkom in oblikuje interval zaupanja. 95% interval za populacijsko vrednost Cohenovega d-ja se razteza od 0.15 do 0.85. Kar naenkrat ni več popolnoma jasno, da je pravi vpliv srednje velikosti, čeprav je bila vzorčna vrednost natančno 0.50. Še več; interval zaupanja razkrije, da je lahko vpliv manjši od majhnega vpliva (to je, manj kot 0.20) ali večji od večjega vpliva (to je, večji od 0.80).

Izvedbo načrtovanja velikosti vzorca lahko pogosto izboljšamo z načrtovanjem velikosti vzorcev, ki vodijo ne do zgolj statistično značilnih, ampak tudi do natančnih ocen parametrov (*angl. accurate parameter estimates*). Gre za t.i. AIPE pristop (*angl. Accuracy in Parameter Estimation*) (Kelley in Maxwell 2003), ki pomeni moderno usmeritev k analizi podatkov.

Namesto da raziskovalec preprosto testira ali je dana ocena parametra neka eksaktna in specificirana vrednost, da oblikovanje 100(1-α)% intervala zaupanja[4] okoli relevantnega parametra pogosto bolj smiselno informacijo. Intervali zaupanja lahko zagotovijo raziskovalcu visoko stopnjo gotovosti, da leži prava vrednost parametra znotraj nekih meja zaupanja. Razumevanje verjetnega obsega vrednosti parametra vodi do boljšega razumevanja preučevanega fenomena kot samo preprosto sklepanje ali je parameter statistično značilen, ali ne. Glede na perspektivo natančnosti ocen parametrov nas čim ožji interval zaupanja vodi v večjo gotovost, da opazovana ocena parametra v večji meri aproksimira ustrezen populacijski parameter (Kelley in Maxwell 2003).

AIPE pristop skuša zmanjšati problem širokih intervalov zaupanja 'ki spravljajo raziskovalce v zadrego'. Pomaga nam oceniti velikosti vzorcev, ki vodijo do dovolj natančnih intervalov zaupanja za smiselno in ne samo statistično značilno oceno. Gre za moderen pristop, kjer intervali zaupanja ter splošna orientacija v meta-analitično razmišljanje vodijo k boljšemu razumevanju raziskovalnih problemov, boljših interpretacij rezultatov in načrta raziskave kot testiranje značilnosti ničelne hipoteze (Kelley in Maxwell 2003).

Načrtovanje velikosti vzorca za Moč nikakor ni nezdružljivo z načrtovanjem velikosti vzorca za Natančnost. Pogosto sta namreč pomembni obe perspektivi in morata biti obravnavani skupaj, saj je v mnogih primerih cilj raziskave, da privede do natančne ocene parametra in da določi, ali je parameter negativen, ničelen, ali pozitiven (Maxwell, Kelley, Rausch 2008).

Intervali zaupanja nam zagotavljajo uporaben organizacijski okvir, da lahko hkrati obravnavamo smer, velikost in natančnost vpliva.

***Smer*** je jasna (znotraj običajnih omejitev verjetnostne gotovosti), ko interval zaupanja ne vključuje ničle. Moč iz te perspektive pogosto izhaja iz želje po zadostno visoki verjetnosti, da interval zaupanja, osnovan na opazovanih podatkih, ne bo vseboval vrednosti 0.

[4] Interval zaupanja, ki ga določata njegova spodnja in njegova zgornja meja, je interval, v katerem se z dano gotovostjo (ponavadi določimo 95 odstotno) nahaja ocenjevani parameter. Interpretacija je naslednja: z verjetnostjo tveganja $\alpha$ se parameter nahaja v tem intervalu.

***Velikost*** zahteva obravnavanje preciznosti in natančnosti. Če je pomembno ocenjevanje velikosti parametra, moramo poleg središča intervala, obravnavati tudi širino intervala zaupanja za ta parameter. Do ozkega intervala privedejo majhne standardne napake ocene parametra, kar je enako trditvi, da je parameter ocenjen precizno.

***Natančnost*** ne vsebuje samo preciznosti, ampak tudi interval, ki naj bi vseboval pravo populacijsko vrednost. V mnogih situacijah, gresta preciznost in natančnost z roko v roki, ker so mnogi cenilci (*angl. estimators*) nepristranski ali vsaj konsistentni. Več o odnosu med natančnostjo in preciznostjo v Kelley in Maxwell 2003, 2008; Kelley in drugi (2003) in Kelley in Rausch (2006).

Načrtovanje velikosti vzorca mora v nekaterih primerih zagotoviti dovolj velik vzorec za zadostno verjetnost zavrnitve ničelne hipoteze, v drugih primerih pa primerno širino intervala zaupanja. Velikost vzorca za natančno oceno je lahko večja od velikosti vzorca, ki jo potrebujemo za zadostno moč, lahko pa je tudi obratno. To je v glavnem odvisno od velikosti vpliva, ki ga želimo izmeriti. V mnogih situacijah je potrebno doseči dva cilja: (a) zavrniti ničelno hipotezo in dokazati smer vpliva (Moč), in (b) oceniti vpliv natančno (*angl. accurately*) (Natančnost). Osnovna ideja načrtovanja velikosti vzorca za natančnost je osnovana na nadzorovanju širine intervala zaupanja (w) našega zanimanja.

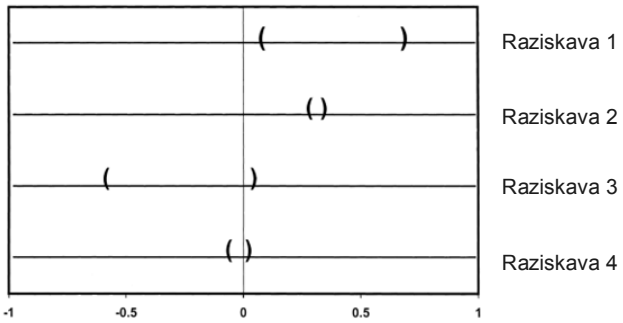
Pristop AIPE zahteva specifikacijo 'tolerance', ki predstavlja verjetnost, da bo interval širši, kot je zaželeno. Npr. raziskovalec želi biti 80% prepričan, da dobi interval, ki ni širši od zaželene vrednosti. V tem primeru bo toleranca enaka 0.20. Takšen cilj jasno zahteva večjo velikost vzorca kot v primeru, da se raziskovalec zadovolji s pričakovano širino intervala. V slednjem primeru zagotovimo takšno velikost vzorca, da je pričakovana širina intervala zaupanja neka vnaprej določena vrednost. Vendar pa bo interval, ki ne bo širši od vnaprej določene vrednosti, realiziran samo v (približno) 50% primerov. Raziskovalec lahko z vnaprej določeno verjetnostjo izračuna velikost vzorca, ki bo zagotovila pričakovano širino intervala. Preciznost intervala zaupanja in stopnja zagotovitve te preciznosti pa sta odvisni od ciljev raziskovalca. Večja preciznost in večja zagotovitev preciznosti predpostavljata večjo velikost vzorca.

## 2.3 AIPE in PA z roko v roki

Znanstvena disciplina plačuje ceno raziskav s premajhno močjo, kljub temu, da jo posamezni raziskovalci nujno ne. Raziskave s premajhno močjo se nagibajo k temu, da proizvajajo literaturo z očitno protislovnimi ugotovitvami. Še več; Goodman in Berlin (1994), Hunter in Schmidt (2004) in Maxwell (2004) so pokazali, da takšna očitna protislovja lahko dejansko ne odražajo nič več kot vzorčno variabilnost. Poročanje rezultatov kot samo statistično značilnih ali neznačilnih še poslabša problem. Veliko bolje bi bilo poročati o rezultatih v smislu intervalov zaupanja, saj prikazujejo negotovost velikosti vplivov in tako preprečujejo bralcem strokovne literature, da bi pretiravali v interpretiranju (*angl. overinterpreting*) prisotnosti mnogih zvezdic poleg majhnih p-vrednosti.

Slika 2.1 predstavlja odnos med intervali zaupanja in testiranjem statistične značilnosti ničelne hipoteze v povezavi s problemom velikosti vzorca za AIPE in PA. Konkretneje, slika prikazuje meje intervala zaupanja za standardiziran regresijski koeficient v štirih hipotetičnih situacijah z različnimi pojasnjevalnimi spremenljivkami v vsakem primeru.

**Slika 2.1: Grafična predstavitev možnih scenarijev v kateri je načrtovana velikost vzorca obravnavana kot 'uspeh' ali 'neuspeh' iz perspektive natančnosti v ocenjevanju parametra in analize moči.**




Iz perspektive analitične moči, raziskavo 1 obravnavamo kot uspeh. Interval zaupanja kaže, da parameter verjetno ni enak nič in je tako ocenjen kot statistično značilen. Vendar pa je interval zaupanja širok in tako parameter ni natančno ocenjen. Imamo

primerno velikost vzorca iz perspektive moči, večji vzorec pa je potreben, če želimo dobiti bolj precizno oceno.

Raziskava 2 kaže, da lahko zavrnemo ničelno hipotezo, poleg tega zagotavlja tudi precizno informacijo o velikosti populacijskega parametra. Tu je interval zaupanja ozek in je populacijski parameter precizno ocenjen. Raziskava dve je obravnavana kot uspeh tako iz stališča PA kot AIPE perspektive.

Raziskava 3 kaže na statistično neznačilen vpliv, ki ga omejuje širok interval zaupanja, kar pomeni neuspeh iz stališča obeh metod. Če bi raziskovalec uporabil večji vzorec in bi bil vpliv približno enake velikosti, bi bila širina intervala zaupanja verjetno ožja, kar bi privedlo do možne zavrnitve ničelne hipoteze.

Raziskava 4 predstavlja primer, kjer interval zaupanja vsebuje 0, vendar pa je parameter ocenjen precizno. Gre za primer neuspeha s stališča PA, a za uspeh s stališča AIPE. Seveda lahko trdimo, da ta raziskava ni dobesedno neuspeh s stališča PA, ker je moč odvisna od populacijske velikosti vpliva. V tej raziskavi je lahko populacijska velikost vpliva manjša od minimalne velikosti vpliva teoretične in praktične pomembnosti.

Cilja PA in AIPE sta v osnovi različna. Cilj PA je dobiti interval zaupanja, ki korektno izključuje ničelno vrednost, kar naredi smer vpliva nedvoumno, jasno. Potrebna velikost vzorca iz te perspektive zavisi od vrednosti vpliva samega. Na drugi strani pa je cilj AIPE dobiti natančno oceno parametra, ne glede na to, ali interval vključuje ničelno vrednost, ali ne. Tako velikost vzorca iz perspektive AIPE ne zavisi od vrednosti vpliva samega. Vendar pa si ti dve metodi načrtovanja velikosti vzorca nista nasprotni; raje nanju gledamo kot na komplementarni.

# 3 Večnivojska analiza: hierarhični linearni modeli

V tem poglavju bom na kratko predstavila večnivojsko analizo in hierarhične linearne modele. Za celostno predstavitev naj se bralec obrne na Bryk in Raudenbush 1992; Goldstein 1995; Hox 2002, Snijders in Bosker 1999.

## 3.1 Uvod v večnivojsko analizo

Večnivojska analiza predstavlja vrsto metod, ki jih uporabljamo pri analizi podatkov s kompleksnimi vzorci variabilnosti, s poudarkom na gnezdenih virih variabilnosti (*angl. nested sources of variability*). Takšne podatke lahko predstavljajo dijaki v razredih, zaposleni v podjetjih, longitudinalna merjenja subjektov, ipd.. Glavna ideja večnivojske analize je, da verjetnostni model *(angl. probability model)* multiple regresijske analize ne predstavlja dovolj natančno večnivojskih podatkov, ki na vseh nivojih vključujejo nepojasnjeno variabilnost.

Večstopenjski vzorci (*angl. multi-stage samples*)[5] se namesto enostavnega slučajnega vzorca uporabljajo zaradi nižjih stroškov, ali pa zato, ker raziskovalca zanimajo odnosi med spremenljivkami na različnih nivojih hierarhičnega sistema. V drugem primeru raziskovalca zanima odvisnost opazovanih enot znotraj skupin, saj se skupine med seboj razlikujejo v določenih pogledih. Uporaba eno-nivojskih statističnih modelov v takšnih primerih ni veljavna.

Raziskovalci pri analizi pogosto zanemarijo dejstvo, da je bila uporabljena večstopenjska shema vzorčenja in tako predpostavljajo, da so bile sekundarne enote izbrane neodvisno. Večstopenjsko vzorčenje namreč vodi k odvisnim opazovanim enotam.[6] Če obravnavamo primer šol in dijakov (npr. dosežki v matematiki dijakov

[5] Če obstaja samo ena 'podpopulacijska' raven, govorimo o dvostopenjskem vzorcu. V slučajnem dvostopenjskem vzorcu v prvem koraku izberemo slučajni vzorec primarnih enot (npr. šole, soseščine, družine) in nato še slučajni vzorec sekundarnih enot (npr. dijaki, družine, otroci), ki jih v drugem koraku vzorčimo iz izbranih primarnih enot.

6 Primarne enote (klastre, enote 2.ravni) bomo od tu dalje imenovali enote na makro nivoju (makro-enote), sekundarne enote (elementarne enote, enote 1.ravni) pa enote na mikro nivoju (mikro-enote). Omejili se bomo na dvonivojske načrte.

znotraj šol), odvisnost lahko izvira iz dejstva, da dijaki znotraj šole (1) živijo v istem šolskem okolju, (2) imajo iste učitelje, (3) vplivajo drug na drugega z neposredno komunikacijo ali skupinskimi normami, (4) izhajajo iz iste soseščine. Bolj ko so dosežki učencev podobni znotraj določene šole (v primerjavi z dijaki iz drugih šol), bolj verjetno je, da na dosežke vpliva organizacijska enota (v tem primeru je to šola).

Pravilno obravnavanje večnivojskih podatkov zaobjema tako odnose znotraj- kot tudi odnose med- skupinami, pri čemer se skupine nanašajo na enote na višjih nivojih gnezdene hierarhije. Pogosto je v takšnem primeru smiselno uporabiti verjetnostne modele, s katerimi predstavimo variabilnost znotraj in med skupinami. Z drugimi besedami, da si predstavljamo nepojasnjena odstopanja med skupinami in nepojasnjena odstopanja znotraj skupin kot slučajno variabilnost. Le-to lahko izrazimo s statističnimi modeli, s t.i. slučajnimi koeficienti. Takšen model slučajnih koeficientov za večnivojske ali hierarhično strukturirane podatke je hierarhični linearni model (v nadaljevanju tudi HLM) in predstavlja glavno orodje za večnivojsko analizo.

Hierarhični linearni model je vrsta regresijskega modela, ki se od navadnega multiplega regresijskega modela razlikuje v tem, da enačba, ki definira HLM vsebuje več kot samo eno napako modela (*angl. error term*) - eno (ali več) za vsak nivo. Kot v vseh regresijskih modelih, obstaja tudi tu razlikovanje med odvisnimi ter pojasnjevalnimi spremenljivkami: cilj je oblikovati model, ki izraža, kako je odvisna spremenljivka odvisna od pojasnjevalnih (neodvisnih) spremenljivk.

Obstaja več vrst uporab hierarhičnih linearnih modelov, ki rangirajo od preprostejših do bolj kompleksnih, vključujoč enostransko analizo variance (ANOVA) s slučajnimi vplivi, regresijski model s povprečji kot odvisno spremenljivko (*angl. means-as-outcomes*), enostranska analiza kovariance (ANCOVA) s slučajnimi vplivi, regresijski model s slučajnimi koeficienti, model s presečišči- in nagibi- kot odvisno spremenljivko in model z ne-slučajno variirajočimi nagibi (*angl. model with nonradomly varynig slopes*) (Raudenbush in Bryk 1993). Vseh podmodelov hierarhičnega linearnega modela ne bomo obravnavali. Pogledali si bomo regresijski model s slučajnimi koeficienti ter enostransko analizo variance s slučajnimi vplivi. Najprej obravnavajmo napreprostejši večnivojski model, t.i. prazen model oz. enostransko analizo variance ANOVA s slučajnimi vplivi.

## 3.2 Enostranska analiza variance ANOVA s slučajnimi vplivi

Model enostranske analize variance ANOVA s slučajnimi vplivi na mikro nivoju označimo kot:

$Y_{ij} = \beta_{0j} + R_{ij}$ **(3.1a)**.

Predpostavljamo, da so vse napake na mikro nivoju $R_{ij}$ normalno porazdeljene z ničelnim povprečjem in konstantno varianco, $\sigma^2$. V tem modelu odvisno spremenljivko znotraj vsake enote na mikro nivoju napovedujemo samo z enim parametrom na makro nivoju, s presečiščem, $\beta_{0j}$. V tem primeru je $\beta_{0j}$ samo povprečna odvisna spremenljivka za j-to skupino (oz. makro enoto). To je, $\beta_{0j} = \mu_{Y_j}$.

Model na makro nivoju predstavlja formula:

$\beta_{0j} = \gamma_{00} + U_{0j}$ **(3.1b)**,

kjer $\gamma_{00}$ predstavlja glavno povprečje (*angl. grand-mean*) odvisne spremenljivke v populaciji, $U_{0j}$ je slučajni vpliv, povezan z j-to makro enoto, pri čemer predpostavljamo, da ima povprečje nič in varianco $\tau_{00}$.

Substitucija enačbe 3.1b v enačbo 3.1a nam da združen model

$Y_{ij} = \gamma_{00} + U_{0j} + R_{ij}$ **(3.2)**,

ki predstavlja model enostranske analize variance z glavnim povprečjem $\gamma_{00}$, z vplivom skupin (z vplivom na makro nivoju) - $U_{0j}$ in z vplivom individualne enote (z vplivom na mikro nivoju) - $R_{ij}$. Gre za model slučajnih vplivov, ker so vplivi skupin analizirani kot slučajni. Skupna varianca $Y_{ij}$ je tako enaka vsoti naslednjih dveh varianc:

$Var(Yij) = Var(U_{0j} + R_{ij}) = \tau_{00} + \sigma^2$ **(3.3)**.

Ocenjevanje modela enostranske analize variance je pogosto uporabno kot preliminarni korak v analizi hierarhičnih podatkov. Na ta način dobimo oceno povprečja (*angl. point estimate*) in interval zaupanja za glavno povprečje, $\gamma_{00}$. Še bolj

pomembno, model zagotavlja informacijo o variabilnosti odvisne spremenljivke na vsakem nivoju. $\tau_{00}$ predstavlja populacijsko varianco med skupinami (variabilnost med skupinami), $\sigma^2$ populacijsko varianco znotraj skupin (variabilnost znotraj skupin). Hierarhični model v enačbah 3.1 in 3.2 je v celoti brezpogojen v tem, da na nobenem nivoju ni določena niti ena pojasnjevalna spremenljivka.

***Koeficient intraklasne korelacije*** Uporaben parameter, povezan z enostransko analizo variance s slučajnimi vplivi, je znotrajrazredni koeficient korelacije *(angl. intraclass correlation coeficient* –v nadaljevanju označen tudi s kratico ICC*),* ki je definiran kot stopnja podobnosti med mikro enotami, ki spadajo v isto makro enoto (meri delež variance odvisne spremenljivke med makro enotami). Opisuje ga naslednja formula:

$$\rho = \frac{\tau_{00}}{(\tau_{00} + \sigma^2)} \qquad (3.4).$$

Ta koeficient meri delež variance pojasnjevalne spremenljivke med skupinami in ga imenujemo tudi vpliv klastra (*angl. cluster effect*). Uporaben je samo v hierarhičnih linearnih modelih, ki vključujejo tudi slučajno presečišče (začetno vrednost).

V nadaljevanju bom opisala različne vrste spremenljivk, ki so lahko prisotne v hierarhičnih linearnih modelih in je njihovo poznavanje pomembno za razumevanje nadaljnjih poglavij.

## 3.3 Spremenljivke v hierarhičnem linearnem modelu

Kot v primeru regresijskih modelov, obstaja tudi v hierarhičnih linearnih modelih razlikovanje med odvisno (odvisnimi) ter neodvisnimi oz. pojasnjevalnimi spremenljivkami. Cilj je oblikovati model, ki izraža, kako je odvisna spremenljivka odvisna od pojasnjevalnih spremenljivk. V najbolj enostavnih hierarhičnih linearnih modelih je vključena le ena odvisna spremenljivka. Le-ta mora biti spremenljivka na mikro nivoju, saj je HLM model za razlaganje nečesa, kar se dogaja na najnižjem, najbolj konkretnem nivoju. Vendar pa je bistvo večnivojske analize prav v tem, da ima odvisna spremenljivka Y v analizah tako individualni (mikro) kot skupinski (makro) vidik. Enako velja tudi za pojasnjevalne spremenljivke na mikro nivoju. Neodvisna spremenljivka X, čeprav gre za spremenljivko na individualnem nivoju, lahko vsebuje tudi vidik skupine. Povprečje neodvisne spremenljivke X v eni skupini je lahko različno od povprečja neodvisne spremenljivke X v drugi skupini. Z drugimi besedami, neodvisna spremenljivka X ima lahko (in pogosto ima) pozitivno varianco med skupinami.

Posebnost hierarhičnih modelov je torej predvsem v tem, da lahko vključimo tudi spremenljivke, ki definirajo samo skupinski vidik. Gre za t.i. spremenljivke na makro nivoju. Nekatere od teh neposredno definirajo skupine (enote na makro nivoju), druge pa prvenstveno opisujejo enote na mikro nivoju (posameznike), vendar so zagregirane na nivo makro enote. Npr., ko se večnivojska struktura nanaša na otroke znotraj družin, je 'vrsta bivališča' spremenljivka, ki je neposredno odvisna od družine, povprečna starost otroka pa je osnovana na agregaciji spremenljivke starost, ki je sama po sebi spremenljivka na mikro nivoju.

Povprečja po skupinah so še posebej pomembna vrsta pojasnjevalne spremenljivke. Povprečje skupine za dano pojasnjevalno spremenljivko na mikro nivoju definiramo kot povprečje vseh posameznikov oz. enot znotraj te skupine. To je lahko pomembna kontekstualna spremenljivka, s katero lahko izrazimo razliko med regresijama znotraj skupin ter med skupinami. Koeficienti obeh vrst regresij se lahko med seboj močno razlikujejo. Regresijski koeficient *znotraj skupine* izraža vpliv, ki ga ima neagregirana pojasnjevalna spremenljivka znotraj skupine, regresijski koeficient *med skupinami*

izraža vpliv agregirane pojasnjevalne spremenljivke na povprečne vrednosti odvisne spremenljivke po posameznih skupinah. Z drugimi besedami, regresijski koeficient med skupinami je v regresijski analizi le koeficient za podatke, ki so agregirani (npr. s povprečenjem) na raven skupine (Snijder in Bosker 1999).

Obstaja pa še ena pomembna vrsta spremenljivke. Gre za t.i. interakcijo med nivoji (*angl. cross-level interaction*). Gre za interakcijo med spremenljivko na mikro nivoju in spremenljivko na makro nivoju, do katere pride zaradi substitucije hierarhičnih linearnih modelov na makro nivoju v model na mikro nivoju.

## 3.4 Model s slučajnimi presečišči in nagibi (*angl. random intercept and slope model*)

V nadaljevanju bomo obravnavali primer, kjer so poleg presečišč slučajni tudi nagibi. Ne samo, da se povprečen uspeh razlikuje po posameznih šolah (slučajno presečišče), ampak se tudi vpliv učenčeve inteligence na uspeh razlikuje od šole do šole (slučajni nagib). V analizi kovariance je takšen pojav znan kot heterogenost regresij po skupinah oz. kot interakcija skupine in sospremenljivke (*angl. group-by-covariate*). V hierarhičnem linearnem modelu ga modeliramo s slučajnimi nagibi.

### 3.4.1 Model z eno pojasnjevalno spremenljivko na mikro nivoju (X)

Najprej si oglejmo večnivojski model v katerem pojasnjujemo odvisno spremenljivko z eno samo spremenljivko na mikro nivoju. Model na mikro nivoju lahko opišemo kot:

$Y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + R_{ij}$ **(3.5)**.

Denimo, da sta presečišče $\beta_{0j}$ in regresijski koeficient oz. nagib $\beta_{1j}$ odvisna od skupine. Na makro nivoju ju lahko razdelimo na povprečni koeficient in od skupine odvisen odklon:

$\beta_{0j} = \gamma_{00} + U_{0j}$ **(3.6a)**

$\beta_{1j} = \gamma_{10} + U_{1j}$ **(3.6b)**, kjer pomeni:

$\gamma_{00}$ povprečno presečišče enot na makro nivoju

$\gamma_{10}$ povprečni regresijski nagib enot na makro nivoju

$U_{0j}$ sprememba presečišča, povezano z j enoto na makro nivoju

$U_{1j}$ sprememba nagiba, povezano z j enoto na makro nivoju

Predpostavljamo, da imajo ostanki na makro nivoju $U_{0j}$ in $U_{1j}$, kot tudi ostanki na mikro nivoju $R_{ij}$, povprečja 0 glede na dane vrednosti pojasnjevalne spremenljivke X.

Substitucija vodi k modelu:

$Y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + U_{0j} + U_{1j} x_{ij} + R_{ij}$ **(3.7)**,

Prvi del enačbe 3.7 $\gamma_{00} + \gamma_{10} x_{ij}$ predstavlja fiksni, $U_{0j} + U_{1j} x_{ij} + R_{ij}$ pa slučajni del, pri čemer $U_{0j} + U_{1j} x_{ij}$ predstavlja vpliv skupine, $\gamma_{10} x_{ij}$ pa regresijski koeficient oz. nagib za skupino j. Člen $U_{1j} x_{ij}$ predstavlja slučajno interakcijo med skupino in X.

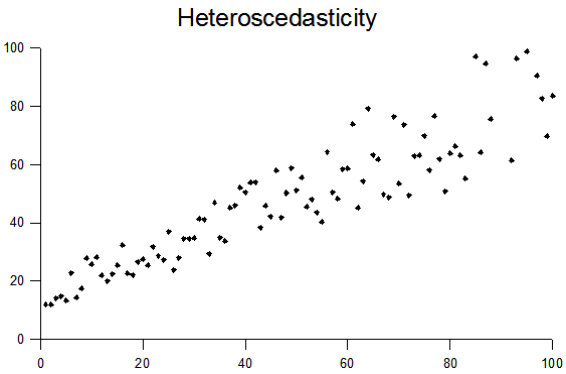
Zgornji model predpostavlja, da skupine označujeta dva slučajna vpliva: presečišče in nagib. Rečemo, da imajo skupine slučajni nagib, slučajni vpliv ali slučajni koeficient. Vpliva ponavadi nista neodvisna, ampak sta med seboj povezana, korelirana. Predpostavlja se, da so za različne skupine pari slučajnih vplivov ($U_{0j}, U_{1j}$) neodvisni in identično porazdeljeni (*angl. independent and identically distributed* i.i.d), da so neodvisni od slučajnih ostankov na mikro nivoju $R_{ij}$ in da so tudi $R_{ij}$ neodvisni in identično porazdeljeni. Varianco $R_{ij}$ označimo kot $\sigma^2$, variance in kovariance slučajnih ostankov na makro nivoju $U_{0j}, U_{1j}$ pa kot:

$\mathrm{var}(U_{0j}) = \tau_{00} = \tau_0^2$ (brezpogojna varianca presečišč na makro nivoju)
$\mathrm{var}(U_{1j}) = \tau_{11} = \tau_1^2$ (brezpogojna varianca nagibov na makro nivoju) **(3.8)**
$\mathrm{cov}(U_{0j}, U_{1j}) = \tau_{01}$ (brezpogojna kovarianca med presečišči ter nagibi na makro nivoju)

Formalno predstavimo pričakovane vrednosti ter razpršenost slučajnih vplivov na makro nivoju (→variančno-kovariančna matrika) takole:

$$E\begin{bmatrix} U_{0j} \\ U_{1j} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad Var\begin{bmatrix} U_{0j} \\ U_{1j} \end{bmatrix} = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix} = T,$$

Model 3.7. predpostavlja, da imajo posamezniki znotraj iste skupine korelirane vrednosti odvisne spremenljivke Y in da je ta korelacija, kot tudi varianca Y, odvisna od vrednosti X. Npr. šole se ne razlikujejo v vplivu socioekonomskega statusa (SES) na uspeh pri učencih z visokim SES, se pa razlikujejo v vplivu pri učencih z nizkim SES. Varianca Y glede na dano vrednost x od X, je odvisna od vrednosti x. Temu pojavu se v statistiki reče heteroskedastičnost.




Regresijska analiza poskuša pojasniti variabilnost odvisne spremenljivke, pri čemer je pojasnjevanje razumljeno v precej omejenem smislu. Regresijska analiza namreč napoveduje vrednost odvisne spremenljivke glede na vrednosti pojasnjevalnih spremenljivk. Nepojasnjena variabilnost v običajni multipli regresijski analizi je tako le varianca ostankov. Variabilnost v večnivojskih podatkih pa ima bolj zapleteno strukturo. V večnivojskem modeliranju je namreč vpletenih več populacij, ena populacija za vsak nivo. Razlaganje variabilnosti v večnivojski strukturi dosežemo z razlaganjem variabilnosti med posamezniki in med skupinami. Če gre za slučajna presečišča in slučajne nagibe na skupinskem nivoju, poskušamo razložiti tako variabilnost presečišč kot tudi nagibov.

### 3.4.2 Model z eno pojasnjevalno spremenljivko na mikro nivoju (X) in eno pojasnjevalno spremenljivko na makro nivoju (Z)

V modelih 3.6 – 3.7 nekaj variabilnosti razlaga regresija X na Y in sicer s členom $\gamma_{10}x_{ij}$ in s slučajnimi koeficienti $U_{0j}, U_{1j}, R_{ij}$, od katerih vsak razlaga različne dele nepojasnjene variabilnosti. Na eni strani lahko poskušamo najti razlago v populaciji posameznikov, to je na mikro nivoju, pri čemer lahko del variance ostankov, označene z $\sigma^2 = \mathrm{var}(R_{ij})$, zmanjšamo z vključitvijo ostalih spremenljivk na mikro nivoju. Glede na to, da se sestava skupin glede na spremenljivke na mikro nivoju lahko razlikuje od skupine do skupine, lahko vključitev takšnih spremenljivk zmanjša varianco ostankov tudi na skupinskem nivoju. Variabilnost lahko razložimo tudi z informacijo, ki jo dobimo iz populacije skupin (iz makro nivoja). Če želimo zmanjšati nepojasnjeno variabilnost povezano z $U_{0j}$ in $U_{1j}$, lahko razširimo enačbe (3.6) tako, da od skupine odvisne regresijske koeficiente $\beta_{0j}$ in $\beta_{1j}$ pojasnjujemo tudi s spremenljivkami na makro nivoju (Z). Z drugimi besedami; regresijski modeli s slučajnimi koeficienti nam omogočajo oceniti variabilnost regresijskih koeficientov

(tako presečišč kot nagibov) na makro nivoju. Naslednji korak predstavlja modeliranje te variabilnosti, npr. z vprašanjem: »*Katere lastnosti šol (enot na makro nivoju) nam pomagajo pojasniti zakaj učenci v nekaterih šolah dosegajo višji povprečni uspeh kot v drugih in zakaj je v nekaterih šolah vpliv socio-ekonomskega statusa na uspeh večji kot v drugih?*«.

Denimo, da imamo takšno spremenljivko Z, pri čemer dobimo regresijske formule za $\beta_{0j}$ in $\beta_{1j}$:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} z_j + U_{0j} \qquad \textbf{(3.9a)}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} z_j + U_{1j} \qquad \textbf{(3.9b)}.$$

$\beta$ tako obravnavamo kot odvisne spremenljivke v regresijskih modelih za populacijo skupin. To so 'latentne regresije', saj $\beta$ ne moremo opazovati brez napake. Substitucija enačb 3.9 v enačbo 3.7 vodi k skupnem modelu

$$\begin{aligned} Y_{ij} &= (\gamma_{00} + \gamma_{01} z_j + U_{0j}) + (\gamma_{10} + \gamma_{11} z_j + U_{1j}) x_{ij} + R_{ij} \\ &= \gamma_{00} + \gamma_{01} z_j + \gamma_{10} x_{ij} + \gamma_{11} z_j x_{ij} + U_{0j} + U_{1j} x_{ij} + R_{ij} \end{aligned} \qquad \textbf{(3.10)}.$$

Zadnji izraz je predelan tako, da imamo najprej 'fiksni' del in nato 'slučajni' del. Če primerjamo ta model s 3.7 se pokaže, da takšna razlaga slučajnih presečišč in nagibov vodi k različnemu fiksnemu delu modela, ne spremeni pa se formula za slučajni del, ki ostaja $U_{0j} + U_{1j} x_{ij} + R_{ij}$. Lahko pričakujemo, da bosta varianci ostankov slučajnega presečišča ter nagiba, $\tau_0^2$ in $\tau_1^2$, manjši od njunih nasprotij v modelu 3.7, ker del variabilnosti presečišč ter nagibov sedaj razlaga tudi spremenljivka Z.

V enačbi 3.10 pojasnjevanje presečišča $\beta_{0j}$ s spremenljivko na makro nivoju (Z) vodi h glavnemu vplivu Z ($\gamma_{01} z_j$), pojasnjevanje koeficienta $\beta_{1j}$ od X, s spremenljivko na makro nivoju Z, vodi k produktu interakcijskega vpliva X in Z ($\gamma_{11} z_j x_{ij}$). Za definicijo interakcijskih spremenljivk kot je produkt $z_j x_{ij}$ v 3.10 je priporočljivo uporabiti transformirane spremenljivke Z in X, tako da imajo vrednosti Z=0 in X=0 nek pomen, ki se ga da interpretirati. Spremenljivki X in Z sta lahko osrediščeni okrog povprečja (*angl. centered around their means*), pri čemer ničelna vrednost spremenljivke pomeni povprečno vrednost. Druga možnost je, da ničelne vrednosti ustrezajo neki referenčni skupini. Razlog za to je, da v prisotnosti interakcijskega člena $\gamma_{11} z_j x_{ij}$

glavni vpliv koeficienta $\gamma_{10}$ od X interpretiramo kot vpliv X za primere, kjer je Z=0, medtem ko glavni vpliv koeficienta $\gamma_{01}$ od Z interpretiramo kot vpliv spremenljivke Z za primere, ko je X=0. Več o osrediščanju spremenljivk v Snijders in Bosker (1999, 80-81) in Raudenbush in Bryk (2002, 31-35).

Interakcije med nivoji lahko obravnavamo na podlagi dveh vrst argumentov. Če raziskovalec ugotovi statistično značilno varianco slučajnih nagibov, lahko sklepa, da obstajajo spremenljivke na makro nivoju, ki bi lahko pojasnjevale ta slučajni nagib. Interakcijo med nivoji pa lahko formuliramo tudi na vsebinsko-teoretičnih argumentih, še preden pogledamo podatke. Na ta način lahko raziskovalec oceni ter testira vpliv interakcije med nivoji ne glede na to, ali je ugotovil varianco slučajnih nagibov ali ne.

### 3.4.3 Splošni model slučajnih presečišč in nagibov

Prejšnji model (3.10) lahko razširimo z vključitvijo večjega števila spremenljivk, ki imajo slučajne vplive in z večjim številom spremenljivk, ki te slučajne vplive razlagajo. Denimo, da imamo p pojasnjevalnih spremenljivk na mikro nivoju, $X_1,...,X_p$ in q pojasnjevalnih spremenljivk na makro nivoju, $Z_1,...,Z_q$. Če raziskovalec nima pomislekov glede uporabe modela s preveč parametri, lahko obravnava model, kjer imajo vse spremenljivke na mikro nivoju (X), variirajoče nagibe in kjer vse spremenljivke na makro nivoju (Z), razlagajo slučajna presečišča kot tudi vse te nagibe.

Na nivoju znotraj skupine predstavimo regresijski model s p-spremenljivkami:

$$Y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + ... + \beta_{pj}x_{pij} + R_{ij} \qquad \textbf{(3.11)}.$$

Razlaga regresijskih koeficientov $\beta_{0j}$ in $\beta_{pj}$ je osnovana na modelu med skupinami, ki je regresijski model s q spremenljivkami za koeficient $\beta_{hj}$:

$$\beta_{hj} = \gamma_{h0} + \gamma_{h1}z_{1j} + ... + \gamma_{hq}z_{qj} + U_{hj} \qquad \textbf{(3.12)}.$$

S substitucijo enačb 3.11 in 3.12 in s prestavitvijo členov dobimo naslednji model:

$$Y_{ij} = \gamma_{00} + \sum_{h=1}^{p}\gamma_{h0}x_{hij} + \sum_{k=1}^{q}\gamma_{0k}z_{kj} + \sum_{k=1}^{q}\sum_{h=1}^{p}\gamma_{hk}z_{kj}x_{hij} + U_{0j} + \sum_{h=1}^{p}U_{hj}x_{hij} + R_{ij} \qquad \textbf{(3.13)}.$$

Na ta način dobimo glavne vplive spremenljivk na mikro in makro nivoju (X in Z) kot tudi vse produkte interakcij med nivoji.

Skupine sedaj označujemo s p+1 slučajnimi koeficienti $U_{0j}$ do $U_{pj}$. Ti slučajni koeficienti so neodvisni med skupinami, lahko pa korelirajo znotraj skupin. Predvidevamo, da je vektor $(U_{0j},...U_{pj})$ neodvisen od ostankov $R_{ij}$ in da imajo vsi ostanki populacijsko povprečje 0, glede na vrednosti vseh pojasnjevalnih spremenljivk. Prav tako se predpostavlja, da so ostanki na mikro nivoju - $R_{ij}$, normalno porazdeljeni, s konstantno varianco $\sigma^2$ in da imajo slučajni vplivi na makro nivoju $(U_{0j},...U_{pj})$ multivariatno normalno porazdelitev s konstantno kovariančno matriko. Podobno kot 3.8 označujemo variance in kovariance slučajnih vplivov na makro nivoju kot:

$$\mathrm{var}(U_{hj}) = \tau_{hh} = \tau_h^2 (h = 1,..., p)$$

$$\mathrm{cov}(U_{hj}, U_{kj}) = \tau_{hk} (h, k = 1,..., p) \qquad \textbf{(3.14)}.$$

## 3.5 Ocenjevanje parametrov

V primeru hierarhičnih linearnih modelov je potrebno razlikovati med *modelom*, ki definira populacijske parametre našega zanimanja, *teorijo ocenjevanja*, ki omogoča, da na vzorčnih podatkih statistično sklepamo o teh parametrih ter *algoritmu računanja*, ki implementira teorijo ocenjevanja. Specifikacija modela vključuje kar nekaj pomembnih izbir: število nivojev v hierarhiji, pojasnjevalne spremenljivke na vsakem nivoju ter najbolj primerno povezovalno funkcijo, ki povezuje pričakovane vrednosti odvisne spremenljivke z množico neodvisnih spremenljivk. Glede na dan model lahko premislimo o možnih pristopih k ocenjevanju. Le-ti vključujejo metodo največjega verjetja (FML) (*angl. full maximum likelihood*), omejeno metodo največjega verjetja (RML) (*angl. residual/restricted maximum likelihood*) in Bayesove metode. Tri metode privedejo do primerljivih rezultatov pri večjih vzorcih, pri majhnih vzorcih pa se lahko rezultati teh metod nekoliko razlikujejo.

V literaturi (npr. Longford, 1993a) navajajo predvsem dve glavni metodi za ocenjevanje statističnih parametrov, pod predpostavko, da so $U_{0j}$, $U_{hj}$ in $R_{ij}$ normalno porazdeljeni: FML in RML. Metodi se med seboj le malo razlikujeta glede ocene regresijskih koeficientov, se pa razlikujeta glede ocenjevanja variančnih komponent. Največja razlika med metodama je, da RML metoda ocenjuje variančne komponente z upoštevanjem izgube (*angl. loss*) prostostnih stopenj, ki izhajajo iz ocenjevanja regresijskih parametrov, FML pa ne. Zaradi tega imajo FML cenilke (*angl. estimators*) za variančne komponente pristranskost s težnjo k padanju (*angl downward bias*), RML cenilke pa ne. Npr. običajna cenilka variance v enonivojskem primeru, kjer RML cenilko predstavlja vsota kvadriranih odklonov deljena z velikostjo vzorca minus 1, FML cenilka pa to vsoto deli še s skupno velikostjo vzorca. Do večjih razlik pride pri manjših velikostih vzorca, medtem ko je pri velikih vzorcih (npr. večjih od 30), ta razlika zanemarljiva. Literatura priporoča, da je RML metodo bolje uporabljati za ocenjevanje parametrov variance in kovariance, testi deviance (*angl. deviance test*s) pa v nekaterih primerih zahtevajo uporabo FML metode (Snijders in Bosker 1999, bolj podroben opis metod v Hox 2002).

Končno, glede na izbor teorije ocenjevanja, potrebujemo algoritem računanja. Popularne izbire vključujejo EM ('expectation-maximization') algoritem (Dempster, Laird in Rubin, 1977; Dempster, Rubin in Tsutakawa, 1981), Fisherjevo metodo (*angl. Fisher scoring*) (Longford, 1987), iterativno posplošeno metodo najmanjših kvadratov (IGLS) (Goldstein, 1986) ter RIGLS (Residual or Restricted IGLS). Gre za iterativne postopke, kar pomeni, da je potrebno določeno število korakov, pri čemer se t.i. provizorične ocene vedno bolj približujejo končni oceni. Po pravilu ti koraki konvergirajo k ML ali RML oceni. Načeloma vsi algoritmi privedejo do enakih ocen za dano metodo ocenjevanja (ML, RML), razlike so pri bolj kompliciranih modelih, v količini komputacijskih problemov ali v času računanja. Izvedba algoritma, glede na težavnost uporabe, variira v smislu mere konvergence in zanesljivosti konvergence (Raudenbush in Bryk 2002, Snijders in Bosker 1999).

## 3.6 Postopki testiranja hipotez

V tem poglavju bomo obravnavali postopke testiranja hipotez. Posebej bomo obravnavali testiranje hipotez za regresijske koeficiente in za komponente variance.

### 3.6.1 Testi za regresijske koeficiente

Denimo, da obravnavamo naslednji model:

$$Y_{ij} = \gamma_0 + \sum_{h=1}^{r} \gamma_h x_{hij} + U_{0j} + \sum_{h=1}^{p} U_{hj} x_{hij} + R_{ij} \qquad (3.15).$$

Ničelno hipotezo, da je **regresijski parameter enak 0**, to je $H_0 : \lambda_h = 0$ , lahko testiramo s t-testom. Statistično ocenjevanje vodi k oceni $\hat{\lambda}_h$ s pripadajočo standardno napako S.E.($\hat{\lambda}_h$). Njuno razmerje je t-vrednost:

$$T(\lambda_h) = \frac{\hat{\lambda}_h}{S.E.(\hat{\lambda}_h)} \qquad (3.16).$$

Na podlagi te testne statistike lahko izvedemo tako enostranske kot dvostranske teste. Gre za enega od pogostih načinov za konstrukcijo t-testa in ga imenujemo Waldov test. Pri veljavni ničelni hipotezi ima $T(\lambda_h)$ približno t-porazdelitev, število prostostnih stopenj (d.f.) je zaradi prisotnosti dveh nivojev nekoliko bolj zapleteno kot v multipli linearni regresiji. Če je skupno število enot na mikro nivoju M, skupno število pojasnjevalnih spremenljivk pa r, potem lahko izračunamo prostostne stopnje kot d.f.=M-r-1. Za testiranje koeficientov na makro nivoju, pri čemer imamo N skupin (enot na makro nivoju) in q pojasnjevalnih spremenljivk na makro nivoju, lahko izračunamo prostostne stopnje kot d.f.=N-q-1. Če je število prostostnih stopenj večje kot 40, lahko t-porazdelitev zamenjamo s standardizirano normalno porazdelitvijo.

### 3.6.2 Test razmerja verjetij (*angl. likelihood ratio test - LR*)

Test razmerja verjetij (znan tudi kot test deviance) predstavlja splošno osnovo za statistično testiranje. Pri uporabi hierarhičnih linearnih modelov ta test večinoma uporabljamo za večparametrične teste (*angl. multiparameter tests*) in za teste slučajnega dela modela.

Ko parametre statističnega modela ocenimo z metodo največjega verjetja, nam ocena da tudi verjetje, ki ga lahko transformiramo v devianco, ki je definirana kot

minus dvakrat naravni algoritem verjetja. To devianco lahko obravnavamo kot mero pomanjkanja prileganja med modelom in podatki. V primeru večine statističnih modelov vrednosti deviance ne moremo interpretirati neposredno, ampak lahko neposredno interpretiramo le razlike v vrednostih deviance za nekaj modelov, ki smo jih testirali na enaki množici podatkov (Snijders in Bosker 1999).

Denimo, da na enaki množici podatkov testiramo dva modela, model $M_0$ z $m_0$ parametri in večji model $M_1$ z $m_1$ parametri. Tako lahko $M_1$ obravnavamo kot razširitev $M_0$, z $m_1$-$m_0$ dodanimi parametri. Denimo, da $M_0$ testiramo kot ničelno hipotezo in $M_1$ kot alternativno hipotezo. Če temu ustrezno označimo devianci z $D_0$ in $D_1$, njuno razliko $D_0$-$D_1$ lahko uporabimo kot testno statistiko, ki ima hi-kvadrat porazdelitev z $m_1$-$m_0$ prostostnimi stopnjami. Ta test lahko uporabimo za parametre fiksnega in slučajnega dela. Vendar pa lahko devianco, ki jo dobimo z metodo RML, uporabimo na ta način samo v primeru, ko imata dva modela ($M_0$ in $M_1$), enak fiksni del in se razlikujeta samo v slučajnih delih.

#### 3.6.2.1 Razpolovljene p-vrednosti v primeru komponent variance

Variance so po definiciji nenegativne. Ko testiramo ničelno hipotezo, da je varianca slučajnega presečišča ali slučajnega nagiba nič, je alternativna hipoteza tako enostranska. To ugotovitev lahko uporabimo za bolj pravilno inačico testa deviance za komponente variance. To načelo sta izpeljala Miller (1977) in Self in Liang (1987). Če je neka varianca enaka 0, potem je verjetnost približno 50%, da je ocenjena vrednost 0 in 50%, da je ocenjena vrednost pozitivna. Če je varianca dejansko ocenjena kot 0, so tudi z njo povezane kovariance ocenjene na 0, s to posledico, da so vsi ocenjeni parametri pod modelom $M_1$ enaki kot pri $M_0$. To vodi k enakosti devianc, $D_1$=$D_0$. Hi-kvadrat porazdelitev drži pod pogojem, da je varianca ocenjena kot neka pozitivna vrednost. Lahko zaključimo, da je ničelna porazdelitev razlike devianc t.i. mešana porazdelitev, z verjetnostjo 50% za vrednost 0 in z verjetnostjo 50% za hi-kvadrat porazdelitev.

Najprej obravnavajmo primer, ko testiramo slučajno presečišče. Ničelni model $M_0$ je model brez slučajnega dela na makro nivoju, kar pomeni, da so vse opazovane enote $Y_{ij}$ neodvisne in pogojne glede na vrednosti pojasnjevalnih spremenljivk. Gre za navaden linearni regresijski model. Alternativni model $M_1$ predstavlja model

slučajnega presečišča z enakimi pojasnjevalnimi spremenljivkami. Obstaja $m_1$-$m_0$ dodatni parameter, varianca slučajnega presečišča. Za opazovane deviance $D_0$ modela $M_0$ (ta model lahko ocenimo z navadno metodo najmanjših kvadratov) in $D_1$ model slučajnega presečišča, izračunamo razliko $D_0$-$D_1$. Če je $D_0$-$D_1$=0, varianca slučajnega presečišča ni statistično značilna (je ocenjena kot 0). Če je $D_0$-$D_1$>0, se slučajnost (*angl. tail probability*) razlike $D_0$-$D_1$ preveri v tabeli hi-kvadrat porazdelitve z d.f.=1. Vrednost p za testiranje statistične značilnosti variance slučajnega presečišča predstavlja polovico ugotovljene verjetnosti.

Sledi primer testiranja slučajnega nagiba. Denimo, da drži model:

$$Y_{ij} = \gamma_0 + \sum_{h=1}^{r} \gamma_h x_{hij} + U_{0j} + \sum_{h=1}^{p} U_{hj} x_{hij} + R_{ij} \qquad \textbf{(3.15)}.$$

Ničelna hipoteza je, da je zadnja varianca slučajnega nagiba nič: $\tau_p^2 = 0$. Pod to hipotezo, so tudi kovariance, $\tau_{hp}$ za h=0,…,p-1, enake 0. Alternativna hipoteza je model, ki ga definira 3.15 in ima $m_1$-$m_0$=p+1 parametrov več kot ničelni model (ena varianca in p kovarianc). Npr., če ni drugih slučajnih nagibov v modelu (p=1), je m1-m0=2. Nadaljujemo z enakim postopkom kot za testiranje slučajnega presečišča: ocenimo oba modela, kar nam da razliko devianc $D_0$-$D_1$. Če je $D_0$-$D_1$=0, varianca slučajnega nagiba ni statistično značilna. Če je $D_0$-$D_1$>0, pa se verjetnost razlike $D_0$-$D_1$ primerja s tabelo hi-kvadrat porazdelitve z d.f.=p+1. P-vrednost za testiranje statistične značilnosti variance slučajnega nagiba predstavlja polovico te vrednosti.

Za bolj podrobno predstavitev naj se bralec obrne na Self in Liang (1987).

# 4 Načrtovanje večnivojske raziskave

V pričujoči magistrski nalogi se bomo osredotočili na specifičen vidik raziskovalnega načrta, konkretneje, na velikosti vzorcev. Vprašanja o velikosti vzorcev v večnivojskih raziskavah so bila obravnavana tudi v Snijders in Bosker (1993), Mok (1995) in Cohen (1998), Raudenbush (1997), Hedeker, Gibbons in Waternaux (1999).

Obstajajo različni načini za izbor velikosti vzorcev v večnivojskih raziskavah, ki glede na dane finančne in druge praktične omejitve zagotavljajo zadostno moč za testiranje oz. majhne standardne napake za ocenjevanje določenih parametrov. Problem v praktični uporabi teh metod je, da velikosti vzorcev, ki so optimalne za testiranje nekega vpliva interakcij med nivoji, niso nujno optimalne za ocenjevanje znotrajrazredne korelacije (naveden samo primer).

Velikost vzorca na najvišjem nivoju je ponavadi najbolj omejevalen element v načrtu. Tako je dvonivojski načrt z desetimi skupinami vsaj toliko neugoden kot enonivojski načrt z desetimi enotami v vzorcu. Pri hierarhičnem linearnem modelu s q pojasnjevalnimi spremenljivkami na makro nivoju so zahteve glede velikosti vzorca na makro nivoju vsaj tako zaostrene kot zahteve glede velikosti vzorca v enonivojskem načrtu s q pojasnjevalnimi spremenljivkami.

Za načrtovanje dobre večnivojske raziskave je priporočljivo najprej določiti primaren cilj raziskave, izraziti ta cilj v testiranem ali ocenjenem parametru in nato izbrati velikosti vzorcev, pri katerih bo ta parameter ocenjen z majhno standardno napako. Ob vsem tem pa je potrebno upoštevati tudi finančne, statistične in druge praktične omejitve. V nekaterih primerih je mogoče preveriti ali bodo velikosti vzorcev zagotovile nizke standardne napake tudi za nekatere druge parametre (sekundarni cilji) (Snijders in Bosker 1999).

## 4.1 Dejavniki, ki vplivajo na učinkovitost raziskovalnega načrta

V večnivojskem modelu raziskovalni načrt vsebuje število enot za vzorčenje na vseh nivojih (omejeno s stroški vzorčenja in praktičnimi omejitvami), poleg vrednosti pojasnjevalnih spremenljivk pa tudi variančno-kovariančno strukturo slučajnih parametrov. Učinkovitost raziskovalnega načrta za ocenjevanje ali testiranje parametrov je odvisna tudi od populacijskih vrednosti parametrov modela. Izračunavanje statistične moči in natančnosti za večnivojske modele je bolj kompleksno kot v enonivojskih načrtih, ker obstajajo tudi številni drugi dejavniki, ki jih moramo v teh izračunih upoštevati, npr. znotrajrazredni koeficient, metoda ocenjevanja, ipd. Le-ti so opisani v nadaljevanju.

### A) Velikost vzorcev

Kot pri navadni regresijski analizi povečanje števila opazovanih enot vpliva na natančnost in na moč tudi v večnivojski raziskavi. Ta problem je bolj kompleksen v analizi večnivojskih podatkov, kjer moramo za natančno oceno neznanih parametrov in dovolj veliko moč testa za testiranje le-teh, vzorčiti zadostno število enot na vseh nivojih. Definirati moramo torej skupno velikost vzorca (M) in velikosti vzorca na mikro (n) in makro nivoju (N).

V večnivojskih modelih moč ni enostavna monotona funkcija velikosti vzorcev na vsakem nivoju pri konstantnih ostalih dejavnikih. V nekaterih primerih (če npr. obstaja velik ICC) enostavno povečanje velikosti vzorca na mikro nivoju ali hkratno povečanje vzorca na mikro in makro nivoju lahko nima dovolj velikega vpliva na moč. V tem smislu je ocenjevanje moči optimizacijski problem, v katerem določimo različne kombinacije velikosti vzorcev na vseh nivojih. Povečevanje velikosti vzorca na najvišjem nivoju (vzorčenje več skupin) je običajno bolj učinkovito kot povečevanje števila posameznikov v skupini.[7]

[7] V primeru, ko raziskovalca zanimajo vplivi spremenljivk na makro nivoju, Snijders in Bosker (1999) za namene zmanjšanja standardnih napak priporočata vzorčenje velikega števila makro enot z relativno malo mikro enot. V primeru vplivov interakcij med nivoji, pri čemer je fiksna samo skupna velikost vzorca (to je c=0), Snijders in Bosker (1993) priporočata vzorčenje največje možne velikosti vzorca makro enot.

Število vzorčenih enot v vsaki posamezni makro enoti je omejeno z omejitvami proračuna in stroški vzorčenja teh enot. Vzorčenje dodatne skupine (brez povečanja skupnega števila enot na mikro nivoju) je ponavadi dražje kot vzorčenje dodatne enote iz skupine, ki je že vključena v raziskavo. Če se npr. odločimo, da vzorčimo dodatnega učenca (na mikro nivoju) znotraj na novo vzorčene šole (na makro nivoju), bi dodatni stroški lahko vključevali kontakte s to šolo in potne stroške. Diferencialni strošek vzorčenih enot na vsakem nivoju tako pripelje do kompromisa med imeti čim več enot na makro nivoju in čim več skupno opazovanih enot (M. Cohen 1998; Mok 1995; Snijders in Bosker 1993).

**B) Stroškovne omejitve**

Raziskovalec ponavadi nima neomejenega proračuna, obisk 3000 naselij in anketiranje ene osebe na naselje pa je veliko dražje kot anketiranje 30 oseb v samo 100 naseljih. V mnogih primerih v praksi je stroškovna funkcija proporcionalna N(n+c), kjer je c razmerje, ki nam pove, koliko večje stroške imamo, če vzorčimo eno dodatno skupino (ne da bi spremenili skupno število vzorčenih enot), kot če vzorčimo eno dodatno mikro enoto znotraj že vzorčene skupine. Pri poštnih in telefonskih anketah z večnivojskimi načrti v večini primerov do takšnega stroškovnega razmerja ne pride, tako da je c=0. Velikosti vzorca na dveh nivojih torej lahko izpeljemo iz naslednje enačbe (Snijders in Bosker 1999):

proračun≥število skupin*(velikost skupin+razmerje stroškov) (**4.1**),

pri čemer je proračun izražen kot število enot, ki bi jih lahko opazovali, če bi vse enote pripadale isti skupini.

**C) Uravnoteženosti enot na makro nivoju**

Uravnoteženost enot v večnivojski raziskavi oz. konkretneje, ali so vse skupine enake velikosti (uravnotežen načrt) ali ne (neuravnotežen načrt), lahko vpliva na rezultate simulacijskih raziskav in predstavlja eno od predpostavk v analitičnih izračunih za različne ocene v večnivojskih raziskavah. Uravnotežen raziskovalni načrt se zdi smiselna strategija zbiranja podatkov, ker ponavadi ni smiselno izbrati več enot na mikro nivoju iz določenih enot na makro nivoju. V praksi pa se stvari pogosto ne izidejo tako enostavno; npr. nekateri učenci izbrani v vzorec so lahko na dan testiranja odsotni, kar vodi v neodgovor. Prav tako je mogoče, da pristopamo k vzorčenju strukturirano; npr. v nekaterih šolah izberemo več dijakov kot v drugih

šolah – morda so nekatere vrste šol redkejše in bi dijake iz takšnih šol radi nadvzorčili.

Stopnja neuravnoteženosti lahko variira na različne načine. Ena polovica od 100 skupin vsebuje po 50 enot, druga polovica pa 150. Večjo neuravnoteženost predstavlja primer, ko imamo le 10 enot v eni polovici skupin in 190 enot v drugi polovici skupin. Obstajajo primeri, ko se le relativno majhen delež skupin značilno razlikuje od velikosti ostalih skupin. Primer je podatkovna struktura z 90 skupinami po 110 enot in z 10 skupinami v velikosti 10, kar nam da povprečno velikost skupine 100.

Rezultati simulacijskih raziskav (npr. Cools, Van den Noortgate in Onghena 2009, Maas in Hox 2004, 2005) so pokazali, da lahko neuravnoteženost zanemarimo v večini primerov, kar nam olajša raziskovanje učinkovitosti in dovoljuje uporabo obstoječih programov. Cools in drugi (2009) omenjajo izjemo za neuravnotežene podatke z veliko večino majhnih skupin. Še več, empirična vzorčna porazdelitev parametrov variance lahko kaže na znatno sploščenost in asimetrijo, kar je odvisno od števila skupin, za slučajni nagib pa je pomemben dejavnik tudi velikost skupine. Pristranskost so našli le v primerih, ko so obravnavali slučajni nagib. Rezultati so pokazali, da na povečanje pristranskosti vpliva premajhno število skupin, pa tudi manjšanje velikosti skupin na račun večjega števila skupin. Te ugotovitve vodijo do sklepa, da za ocenjevanje variance slučajnega nagiba dodajanje števila majhnih skupin ni zelo informativno. Za kovarianco s takšnim slučajnim nagibom velja enako pravilo. Za ocenjevanje varianc slučajnih presečišč ter regresijskih koeficientov se obrestuje dodajanje majhnih skupin. Zaključki za statistično testiranje so v skladu z zaključki za ocenjevanje. Če vzorčimo dodatne majhne skupine, ne pridobimo veliko moči za ocenjevanje variance slučajnega nagiba.

### E) Znotrajrazredna korelacija

Podatki o odvisni spremenljivki znotraj katerekoli dane makro enote niso statistično neodvisni v hierarhično gnezdenih podatkih. Z drugimi besedami, obstaja povezanost znotraj posamezne skupine, zato je potrebno vzeti v obzir mero povezanosti med podatki, kot je znotrajrazredna korelacija.

## F) Sospremenljivke na mikro in makro nivoju

Vključevanje sospremenljivk na mikro in makro nivoju lahko vpliva na statistično moč in natančnost v večnivojskih modelih (npr. Raudenbush 1997). Sospremenljivke lahko zmanjšajo varianco med skupinami in temu ustrezno spremenijo optimalno alokacijo velikosti vzorca na vsakem nivoju. Bolj konkretno, zaradi vključitve sospremenljivk lahko vzorčimo manj skupin in več posameznikov v teh skupinah (Reise in Duan 2003).

## G) Metoda ocenjevanja

V večnivojskih raziskavah konkretna metoda ocenjevanja lahko vpliva na statistično moč in potrebno velikost vzorca. Na tej točki je le malo priporočil glede metode ocenjevanja (Reise in Duan 2003). Večina večnivojskih raziskav uporablja metodo RML ali FML, pri čemer RML običajno privede do boljših rezultatov kot FML (npr. Maas in Hox 2004).[8]

[8] Formule, ki jih predstavljamo v naslednjem poglavju predpostavljajo uporabo RML postopkov ocenjevanja.

# 5 Pristopi k načrtovanju večnivojske raziskave

Obstajajo tako analitični kot simulacijski pristopi k raziskovanju učinkovitosti načrta v večnivojskih raziskavah. Analitični pristopi temeljijo na rabi formul, iz katerih je moč izpeljati standardne napake in moč testa za ocenjevanje in testiranje specifičnih koeficientov določenega večnivojskega modela. Simulacijski pristopi predstavljajo uporabo simulacij za specifičen raziskovalni načrt. Oba pristopa imata svoje omejitve. Analitične raziskave tako omogočajo le obravnavo preprostih modelov in lahko obravnavajo/manipulirajo le nekaj predpostavk, rezultate simulacijskih raziskav pa je težko posploševati, saj so podatki generirani za specifičen model in za specifične vrednosti parametrov (Cools in drugi 2008).

## 5.1 Analitična metodologija

Analitična metodologija zadeva izpeljavo formul za različne komponente večnivojskih modelov.

### 5.1.1 Ocenjevanje (skupnega) populacijskega povprečja

Najbolj preprost primer večnivojske raziskave predstavlja ocena populacijskega povprečja za določeno relevantno spremenljivko (npr. dohodek, starost, pismenost), pri čemer anketiranci pripadajo različnim regijam. Cochran (1977, 9.pogl.) predstavi formule za izračun zaželene velikosti vzorcev v primeru dvostopenjskega vzorčenja. Definira t.i. vzorčni učinek (*angl. design effect*), ki vsebuje faktor, za katerega se poveča varianca ocene (ki je kvadrirana standardna napaka te ocene) zaradi uporabe dvostopenjskega vzorca namesto enostavnega slučajnega vzorca iste skupne velikosti. Vzorčni učinek tako ponazarja formula:

$$deff = (1 + (n_{skupina} - 1)\rho) \qquad (5.1),$$

kjer n predstavlja povprečno velikost vzorca skupin, ρ pa je znotrajrazredni korelacijski koeficient.

### 5.1.2 Ocenjevanje povezav med spremenljivkami

Do leta 1993, ko sta avtorja Snijders in Bosker objavila pionirski članek z naslovom *»Standard errors and sample sizes for two-level research«* je bilo le malo znanega o tem kako naj bi raziskovalci izbrali velikost vzorca na makro in mikro nivoju, da bi lahko zagotovili zaželeno stopnjo moči glede na relevantno velikost vpliva (vpliv v hipotezi) in izbrano stopnjo statistične značilnosti (α) (Snijders in Bosker 1993). Avtorja sta v članku raziskovala izpeljave standardnih napak pri velikih vzorcih ter statistično moč regresijskih koeficientov v raziskovalnih načrtih z vplivi spremenljivk na makro nivoju ter z vplivi interakcijskih spremenljivk na odvisno spremenljivko. Posameznih formul zaradi večje kompleksnosti in nerelevantnosti v pričujoči magistrski nalogi ne bom posebej obravnavala, bralec lahko več informacij najde v zgoraj opisanem članku ali v priročniku za program PinT (Snijders in Bosker 1993).

Obstajajo tudi druge, manj zapletene formulacije, ki izhajajo iz aproksimacij izračunov enostavnih fiksnih vplivov v večnivojskih modelih in so jih uporabili v celi vrsti raziskovalnih ter vzorčnih načrtov (npr. Bryk in Raudenbush 1992, Cohen 1998, Longford 1993, Snijders in Bosker 1993). Poleg tega, da so manj zapletene, vključujejo tudi manj parametrov, ki jih je potrebno vstaviti v formulo, kar lahko močno olajša delo raziskovalcu. Formule predpostavljajo uravnotežen raziskovalni načrt (enake velikosti vzorcev v vsaki skupini). Ker večina primerov v praksi ne vključuje takšne uravnoteženosti, lahko raziskovalec v tem primeru uporabi minimalno ali povprečno velikost vzorca znotraj vsake skupine.

Raudenbush (1997) je predstavil formulo za **ocenjevanje variance člena nagiba za spremenljivko na makro nivoju** ($\gamma_{01}$)[9] (*glej Enačba 3.9a, 3.10*)

$$\mathrm{var}(\gamma_{01}) = \frac{4(\tau_{00} + \sigma^2 / n)}{N} \qquad (5.2),$$

pri čemer $\tau_{00}$ predstavlja varianco presečišča, $\sigma^2$ pa individualno varianco ostankov (na mikro nivoju).

[9] Glede na to, da je statistična moč povezana s testi za presečišče le redko področje zanimanja raziskovalcev (Scherbaum in Ferreter 2007), predstavljamo samo enačbe za ocenjevanje standardnih napak za nagib v enačbi 3.10.

Enačbo 5.2 lahko napišemo tudi kot (glej Raudenbush in drugi 2005):

$$\mathrm{var}(\gamma_{01}) = \frac{4(\rho + (1-\rho)/n)}{N} \qquad (5.3),$$

kjer zgornji varianci ustrezno zamenjamo s formulo za znotrajrazredni korelacijski koeficient ρ. Če je v model vključena tudi sospremenljivka na mikro nivoju, je formula za varianco nagiba nekoliko spremenjena (glej Raudenbush 1997):

$$\mathrm{var}(\gamma_{01}|X) = \frac{4(\rho_{y|x} + (1-\rho_{y|x})/n)}{N}(1 + \frac{1}{JN-4}) \qquad (5.4).$$

Standardne napake seveda izračunamo s kvadratnim korenom teh varianc. Potrebno se je zavedati, da bo kvaliteta te ocene samo toliko dobra kot je dobra kvaliteta naših ocen za populacijske parametre v enačbi.

### 5.1.3 Komponente variance

Do sedaj so nas zanimali primeri, kjer smo želeli oceniti povprečno vrednost spremenljivke ali povezavo med dvema spremenljivkama. Dodatna zanimiva vprašanja pa se lahko nanašajo na t.i. slučajne vplive. Ali je pomembno katero šolo obiskuje učenec? Ali se soseske razlikujejo v njihovi religiozni sestavi? Ali se otroci razlikujejo v vzorcih rasti? Ali je vpliv socialnega statusa na dohodek različen v različnih regijah? V vseh teh primerih nas zanima velikost komponent variance. Na moč testov komponent variance vplivajo število skupin, velikost skupin in skupna velikost vzorca na drugačen način kot na moč testov za regresijske koeficiente.

Za dvonivojski model s slučajnim presečiščem je enačbe za ocenjevanje variance na mikro nivoju $\sigma^2$ in variance presečišča $\tau_{00}$ izpeljal Longford (1993a, p.58). Longfordovo **enačbo za varianco presečišča** (oz. standardno napako) lahko napišemo kot:

$$\mathrm{var}(\tau_{00}) \approx \frac{2\sigma^4}{nN}(\frac{1}{n-1} + 2\left(\frac{\tau_{00}}{\sigma^2}\right) + n\left(\frac{\tau_{00}}{\sigma^2}\right)^2) \qquad (5.5a) \text{ oz.}$$

$$SE(\tau_{00}) \approx \sigma^2 \frac{2}{Nn}\sqrt{\frac{1}{n-1} + \frac{2\tau_{00}}{\sigma^2} + \frac{n\tau_{00}^{\ 2}}{\sigma^4}} \qquad (5.5b),$$

pri čemer je σ² ocena variance znotraj skupin, $\tau_{00}$ ocena variance med skupinami za presečišče.

Longfordova **formula za ocenjevanje variance** (oz. standardne napake) **na mikro nivoju** je:

$$\mathrm{var}(\hat{\sigma}^2) \approx \frac{2\sigma^4}{nN - N} \qquad \textbf{(5.6a)}$$ oz.

$$SE(\hat{\sigma}^2) \approx \sigma^2 \sqrt{\frac{2}{nN - N}} \qquad \textbf{(5.6b)},$$

kjer so N, n ter σ² enaki kot zgoraj.

Tudi v primeru večnivojskih raziskav statistično moč za testiranje komponente variance na mikro nivoju maksimiziramo z večjimi velikostmi vzorca na mikro nivoju. Minimalno standardno napako tega koeficienta uporabimo samo v primeru, ko nas ta ocena zares zanima, saj je za dosego dobrega raziskovalnega načrta za ocenjevanje parametrov variance bolj pomembna varianca presečišča (M. Cohen 1998, Snijders in Bosker 1999). Ko določamo optimalno velikost vzorca za ocenjevanje variance presečišča, lahko standardne napake prikažemo v grafu z n, s skupno velikostjo vzorca M=N*n, omejeno na dano vrednost, za različne vrednosti znotrajrazrednih koeficientov. Če obstaja omejitev proračuna, lahko v zgornje izraze za Longfordove aproksimacije substituiramo N s K/(n+c) in v grafu prikažemo standardne napake, pri čemer ohranjamo fiksen N*(n+c). Za prikaz primera naj se bralec obrne na Cohena (Cohen 1998, 5.pogl.).

**Standardne napake varianc slučajnih nagibov** predstavljajo precej zapletene funkcije velikosti vzorcev, odstopanj pojasnjevalnih spremenljivk znotraj in med skupinami in samih varianc. Potrebne so nadaljnje raziskave za izpeljavo navodil za izbor primernih velikosti vzorca za ocenjevanje parametrov variance v modelih s slučajnimi nagibi. Če je potrebna precizna ocena takšnih parametrov, je smisleno uporabiti večje vzorce (30 ali več) na obeh nivojih in se prepričati, da je pojasnjevalna spremenljivka, za katero želimo oceniti slučajni nagib, dovolj razpršena znotraj skupin. To lahko dosežemo npr. z neko vrsto stratifikacije (Snijders in Bosker 1999). Za kompleksne modele, kjer niso na voljo formule za standardne napake, lahko dobimo vtis o standardnih napakah z Monte Carlo simulacijo.

## 5.2 Simulacijske raziskave o vplivih na natančnost ter moč v večnivojskih modelih

Obstaja kar nekaj simulacijskih raziskav o vplivih na natančnost in moč v večnivojskih modelih, ki se v večini ukvarjajo z natančnostjo ocen regresijskih koeficientov in komponent variance z majhnimi vzorci tako na makro kot na mikro nivoju. Precej manj raziskav je narejenih na temo natančnosti standardnih napak. Večina simulacijskih raziskav, ki se nanašajo na natančnost testov statistične značilnosti ali pokritosti intervala zaupanja je osnovanih na asimptotičnem razmišljanju: standardna napaka se uporablja v kombinaciji s standardno normalno porazdelitvijo, pri čemer se računa p-vrednosti ali interval zaupanja.[10] Te simulacijske raziskave predstavljajo neka splošna pravila za priporočene velikosti vzorcev, vendar pa gre za vsebino, ki se še razvija. Raziskovalec se ne sme v celoti zanašati na rezultate simulacijskih študij, saj le-te vključujejo vrsto specifičnih predpostavk in ne vključujejo primernih velikosti vzorcev, stopenj moči ali vključitev proračunskih omejitev (Scherbaum in Ferreter 2008). V nadaljevanju bom povzela nekaj glavnih ugotovitev objavljenih simulacijskih raziskav.

***Regresijski koeficienti in standardne napake*** Obstaja kar nekaj simulacijskih raziskav, ki za veliko število velikosti vzorcev na obeh nivojih, obravnavajo standardne napake za enostavne regresijske koeficiente, različne velikosti vplivov ter različne znotrajrazredne korelacijske koeficiente. Bassiri (1988), Browne in Daper (2000), Kim (1990) in Mok (1995) so ugotovili, da ima v primeru regresijskih koeficientov večanje števila skupin večji vpliv na povečanje moči kot povečanje velikosti skupin in to v primeru večjega razpona znotrajrazrednega korelacijskega koeficienta.

Simulacijske raziskave o natančnosti regresijskih koeficientov in njihovih standardnih napak so potrdile, da so ocene za regresijske koeficiente v splošnem nepristranske,

[10] Obstajajo še drugi pristopi, ki bi lahko bili pri majhnih vzorcih bolj veljavni, ampak niso dostopni v vseh večnivojskih programih. Testiranje varianc z uporabo njihovih standardnih napak ni optimalno, ker predvideva normalnost in ker ničelna hipoteza, da je varianca nič, predstavlja test, ki meji na dovoljen prostor parametrov (variance ne morejo biti negativne), kjer standardna teorija verjetja (*angl. likelihood theory*) ni več verjetna. Veliko alternativnih pristopov je bilo predlaganih (cf. Berkhof in Snijders, 2001 za pregled).

tako za OLS (Ordinary Least Squares), GLS (Generalized Least Squares) ter ML (Maximum Likelihood) metodo ocenjevanja (Van der Leeden in Busing 1994; Van der Leeden, Busing in Meijer 1997). OLS ocene so manj učinkovite; Kreft (1996), ki je ponovno analiziral rezultate Kim (1990) je ugotovil, da so OLS ocene le približno 90% učinkovite. Van der Leeden in drugi (1997) so s svojo simulacijsko študijo ugotovili, da imajo standardne napake regresijskih koeficientov, osnovane na ML, kljub temu da ni zadoščeno predpostavkam o normalnosti in velikih vzorcih, le majhno pristranskost navzdol. V splošnem se zdi večje število skupin bolj pomembno kot število posameznikov na skupino.

V novejših raziskavah (npr. Maas in Hox 2004, 2005) so preverjali natančnost standardnih napak regresijskih koeficientov in komponent varianc za različne velikosti vzorcev na obeh nivojih in različne znotrajrazredne korelacijske koeficiente. Natančnost je pomembna, ker lahko standardne napake, ki so pristranske bodisi navzgor ali navzdol, pre- ali pod- cenijo moč in zahtevane velikosti vzorcev. Maas in Hox (2004,2005) sta ugotovila, da ima velikost vzorca na makro nivoju, ki je večja od 30, minimalen vpliv na natančnost standardnih napak za regresijske koeficiente. Velikost vzorca, ki je manjša od 30, pa vodi do premajnih standardnih napak, posebej v primeru večjega znotrajrazrednega korelacijskega koeficienta. Ugotovila sta tudi, da velikost znotrajrazrednega koeficienta v večji meri kot na standardne napake enostavnih fiksnih vplivov vpliva na standardne napake komponent variance.

***Interakcije med nivoji*** V simulacijskih raziskavah Bassiri (1988) in van der Leeden in Busing (1994) so raziskovali moč testa za interakcije med nivoji pri različnih velikostih vzorcev na obeh nivojih. Kot z enostavnimi fiksnimi vplivi, so v obeh raziskavah ugotovili, da ima povečevanje velikosti vzorca na makro nivoju, večji vpliv na moč kot povečevanje velikosti vzorca na mikro nivoju. Ti rezultati so v skladu s simulacijo Snijdersa in Boskerja (1993), ki sta ugotovila, da ima velikost vzorca na makro nivoju večji vpliv na ocene statistične moči kot velikost vzorca na mikro nivoju. Za 'potrditev' interakcij med nivoji, se Busing (1993) kot tudi van der Leeden in Busing (1994) zavzemajo za minimum 30 skupin s po 30 enotami znotraj vsake skupine, da bi dosegli zadostno moč. Če ima nekdo večje velikosti vzorcev na makro nivoju, manjše velikosti vzorcev na mikro nivoju lahko še vedno privedejo do visokih nivojev statistične moči.

***Komponente variance in standardne napake*** Ocene variance na mikro nivoju so ponavadi zelo natančne. Variančne komponente na makro nivoju pa so ponekod podcenjene. Simulacijske raziskave Busing (1993) in Van der Leeden in Busing (1994) kažejo na to, da za natančne ocene varianc na skupinskem nivoju potrebujemo veliko skupin (več kot 100) (cf. Afshartous, 1995).

Simulacije Van der Leeden in drugih (1997) kažejo na to, da so standardne napake, ki jih uporabljamo za testiranje variančnih komponent v splošnem ocenjene kot premajhne, pri čemer se je RML metoda izkazala za bolj natančno metodo kot FML. Simetrični intervali zaupanja okoli ocenjene vrednosti prav tako ne delujejo najbolje. Browne in Draper (2000) poročata podobne rezultate. Tipično, s 24 do 30 skupin, Browne in Draper poročata operativno stopnjo alfa okoli 9%, z 48 do 50 skupin okoli 8%.

S simulacijskimi raziskavami so Afshartous (1995), Busing (1993) in Van der Leeden in Busing (1994) ugotovili, da je za velikosti vzorcev na makro nivoju, ki so manjše od 300, standardna napaka za komponente variance na makro nivoju, podcenjena. V tem primeru bo statistična moč precenjena in primerni vzorci podcenjeni. Maas in Hox (2004,2005) sta ugotovila, da so z velikostmi vzorcev na makro nivoju, velikosti 30 ali 50, standardne napake za komponente variance premajhne, kar vodi k precenjenosti moči in podcenjenosti velikosti vzorca. Če nas torej zanimajo komponente variance, moramo vzorčiti okoli 100 skupin. Ponovno se zdi večje število skupin bolj pomembno kot večje število posameznikov na skupino.

### 5.2.1 Povzetek prejšnjih simulacijskih raziskav

Iz prejšnjih raziskav lahko strnemo nekaj splošnih sklepov glede velikosti vzorcev in statistične moči. Obstaja kompromis med številom enot na mikro in makro nivoju. V splošnem večjo statistično moč dosežemo z večjimi vzorci na makro nivoju kot na mikro nivoju. Ocene za komponente variance na mikro nivoju se zdijo edini parameter za katerega potrebujemo večje velikosti vzorcev na mikro nivoju. Osnovano na pregledu simulacijskih raziskav, Kreft (1996) predlaga pravilo 30/30 za vse vrste vplivov. Bolj konkretno, zagovarja minimum 30 skupin z vsaj 30 enotami v skupini, kar vodi v skupni vzorec 900 enot. Hox (1998) zagovarja še večjo skupno

velikost vzorca s pravilom 50/20. Za ocene, ki zadevajo presečišče, so možne manjše velikosti vzorcev, saj je presečišče ocenjeno bolj precizno kot nagibi (Hofmann 1997). Za ocene, ki vključujejo komponente variance na makro nivoju, so potrebne velikosti vzorcev na makro nivoju, ki so večje od 30 (Maas in Hox 2004, 2005).

Rezultati simulacijskih raziskav predstavljajo splošna vodila in jih nikakor ne moremo zamenjati za izvedbo analize moči. Rezultati številnih simulacijskih raziskav predstavljajo različne nasvete o minimalnem številu skupin ali velikosti teh skupin. Prav tako vodila niso praktična za raziskave, kjer je nemogoče doseči takšno velikost vzorca. V mnogih primerih lahko tudi velikosti vzorcev, ki so mnogo manjše od 900, vodijo do zadovoljive statistične moči, ko nas npr. zanimajo srednji ali veliki vplivi. Moč v večnivojskih modelih ni linearna funkcija in tako dodatna skupina lahko nima bistvenega vpliva na stopnjo moči. Še več, pogosto se zgodi, da so makro enote izhodišče našega teoretičnega zanimanja in jih je v raziskavi na voljo manj kot 30. V teh primerih je potrebna različna razvrstitev (*angl. allocation*) enot med nivoji, s katero uravnotežimo moč in teoretične zahteve. Glede na to, da so z vzorčenjem na različnih nivojih analize povezani tako posredni kot neposredni stroški, je priporočljivo dejansko oceniti moč in zahtevane velikosti vzorcev, saj le tako lahko zagotovimo učinkovito uporabo sredstev in zaželenih verjetnosti za ugotavljanje vplivov, ko le-ti obstajajo (Scherbaum in Ferreter 2008).

# 6 Programi za določanje učinkovitih načrtov večnivojskih raziskav

Za analitične izračune, ki sem jih predstavila v poglavju 5.1. lahko uporabimo enostavne programe kot je Excell, vendar pa izračunavanje teh enačb lahko hitro postane zelo neučinkovito. Na voljo so specializirani programi za izračunavanje moči, standardnih napak in potrebnih velikosti vzorcev v večnivojskih modelih. Le-te bom na kratko opisala v nadaljevanju. Obstajajo tako programi, ki se poslužujejo analitične metodologije, kot tudi novejši programi, ki se poslužujejo simulacijske metodologije.

## 6.1 Programi, ki se poslužujejo analitične metodologije

Le nekaj raziskav se je dejansko končalo s programom, ki omogoča avtomatično izpeljavo analitičnih izračunov za nekatere parametre večnivojske raziskave (npr. PINt - Power in Two-level designs (Snijders in Bosker 1993), OD - Optimal Design (Raudenbush in Liu 2001)). V nadaljevanju bom ta dva programa tudi na kratko opisala.

### 6.1.1 PINt – the power in two-level design

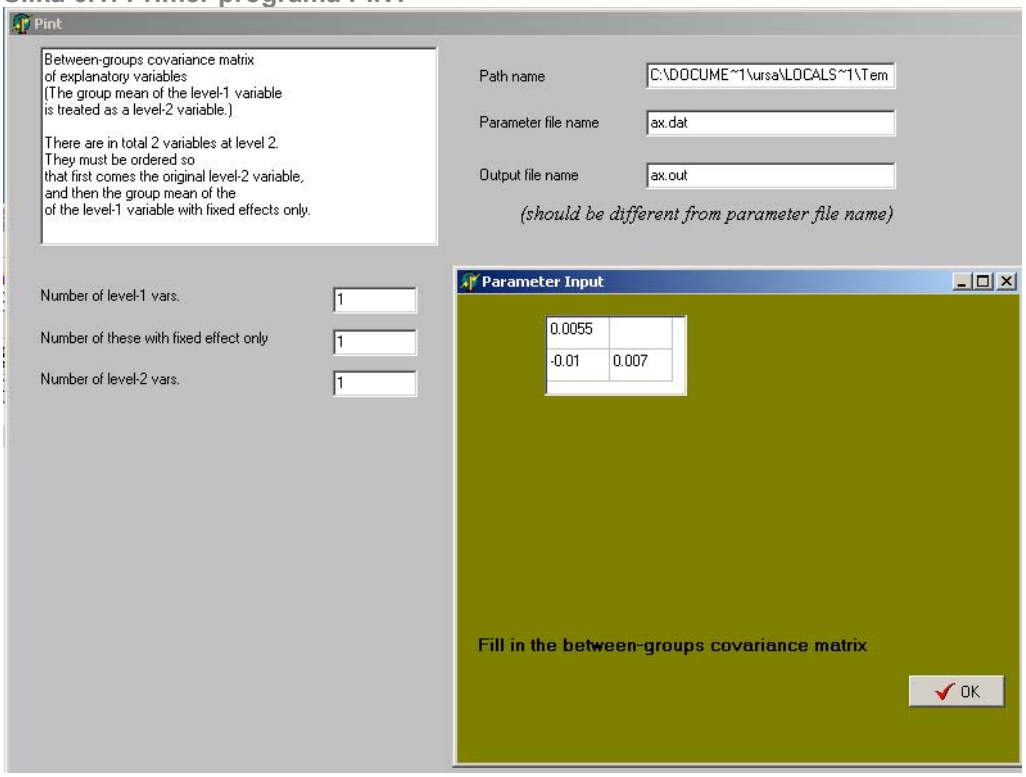
Gre za najstarejši program, ki so ga razvili Snijders in drugi (Snijders in Bosker 1993; Snijders, Bosker in Guldemond 1996)[11] in je namenjen izračunavanju standardnih napak ocen regresijskih koeficientov (tudi interakcij med nivoji) v hierarhičnih linearnih modelih. Dobra lastnost programa je, da so uporabljene formule predstavljene v Snijders in Bosker (1993), na voljo pa je tudi obširen priročnik za uporabnike.

Program izračuna ocene standardnih napak za številne kompleksne modele. Program se priporoča za modele, ki vključujejo več spremenljivk na mikro ter makro nivoju. Velika težava pri uporabi tega programa je, da mora uporabnik vnesti celo vrsto parametrov. Tako potrebujemo informacijo o povprečjih, variancah in

[11] Program PINT in priročnik sta na voljo zastonj na strani: http://stat.gamma.rug.nl/snijders/multilevel.htm

kovariancah pojasnjevalnih spremenljivk, kot tudi o variancah ter kovariancah slučajnih vplivov. Če želimo standardne napake prevesti v statistično moč, moramo oceniti tudi velikost vpliva. PINt tudi avtomatično izpelje pogoje za primerjavo, glede na proračun in stroške vzorčenih enot na vsakem od dveh nivojev.

Slika 6.1: Primer programa PINT




Kot trdita Snijders in Bosker (1993), programa ne moremo uporabiti za izpeljavo standardnih napak za komponente varianc. Nadalje, če imamo zgolj nekaj opazovanih enot ali manj skupin, so lahko aproksimacije pristranske (Snijders in Bosker 1993). V grobem naj bi raziskovalec določil vsaj 6 enot v 10 skupinah. Poleg slabosti, da je program PINt osnovan na aproksimacijah velikih vzorcev, je primeren samo za uporabo dvonivojskih uravnoteženih raziskovalnih načrtov.

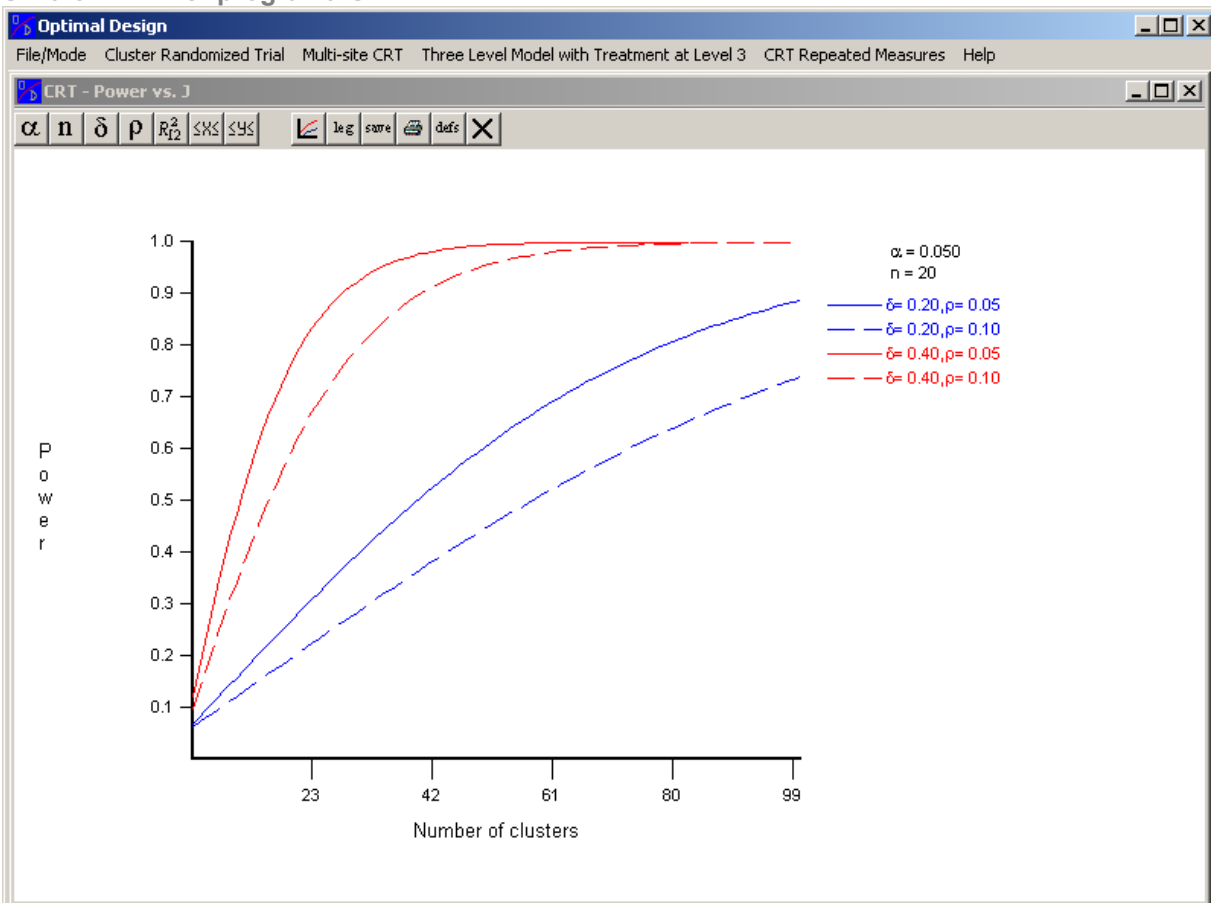
### 6.1.2 OD – optimal design

Raudenbush in sodelavci (2005) so razvili program, ki so ga poimenovali Optimal Design, ki ocenjuje moč z uporabo znotrajrazredne korelacije, velikosti vpliva, stopnje značilnosti α in velikosti vzorcev za posebne večnivojske načrte.[12] Uporabnik lahko posebej spreminja po en dejavnik in tako raziskuje vpliv na moč. Vse rezultate predstavi grafično kot krivulje moči, tako da postane jasno razvidno kako na moč

[12] Program OD ter priročnik sta na voljo zastonj na strani: http://sitemaker.umich.edu/group-based.

vplivajo konkretne spremembe velikosti vzorcev, velikosti vplivov in znotrajrazredne korelacije. Gre za uporabnikom prijazen program in tudi k temu je pridan obširen priročnik.

Slika 6.2: Primer programa OD




Glede na to, da je program osnovan na analitičnih izpeljavah, ki se zanašajo na omejujoče predpostavke, pa zanj veljajo podobne omejitve kot za program PINt.

Obe orodji lahko uporabljamo za raziskovanje osnovnih dinamik med večnivojsko strukturo podatkov in učinkovitostjo, obravnavano iz vidika preciznosti (PINt) ali moči (OD), kar nam omogoča zelo hiter vpogled v problem optimalne alokacije pri večnivojski analizi podatkov. Še več, ker je ocenjevanje regresijskih koeficientov nepristransko, ko so vključeni dovolj veliki vzorci na višjih nivojih (Maas in Hox 2005), je preciznost enaka natančnosti. Za ocene varianc obstaja pristranskost, predpostavka o normalnosti pa je ponavadi nerealistična.

Če bi primerjali programa z istimi vstopnimi parametri, bi zelo verjetno prišlo do razlik v ocenah in optimalnih velikosti vzorcev, vendar so te razlike običajno majhne in ne pripeljejo do nasprotujočih zaključkov.

### 6.1.3 Drugi programi

Rose in Bowen (2005) sta napisala vrsto SAS makrov, s katerimi lahko izračunamo moč za dvo- in tri-nivojske modele za longitudinalne načrte.[13] Brown in Liao (1999) opisujeta spletno orodje za izračunavanje moči v večnivojskih longitudinalnih načrtih.[14]

Na žalost pa mnogih pomembnih situacij ne moremo obravnavati analitično, ker bi sprostitev nekaterih predpostavk takoj otežilo izpeljavo formul – npr. ko je na makro nivoju vzorčenih le malo enot ali ko so vključeni dodatni nivoji. Še več, izpeljave so v glavnem omejene na fiksni del večnivojskega modela, medtem ko lahko raziskovalca primarno zanima prav slučajni del. Za takšne raziskovalne situacije je izpeljava zahtevanih formul v zaprti obliki celo nemogoča, kar vodi k nujni zamenjavi s simulacijskimi postopki.

[13] SAS makre lahko prenesemo zastonj iz: http://www.schoolsuccessprofile.org/analyticaltools.asp

[14] http://psmg.usf.edu/PowerCal/PowerCal.html

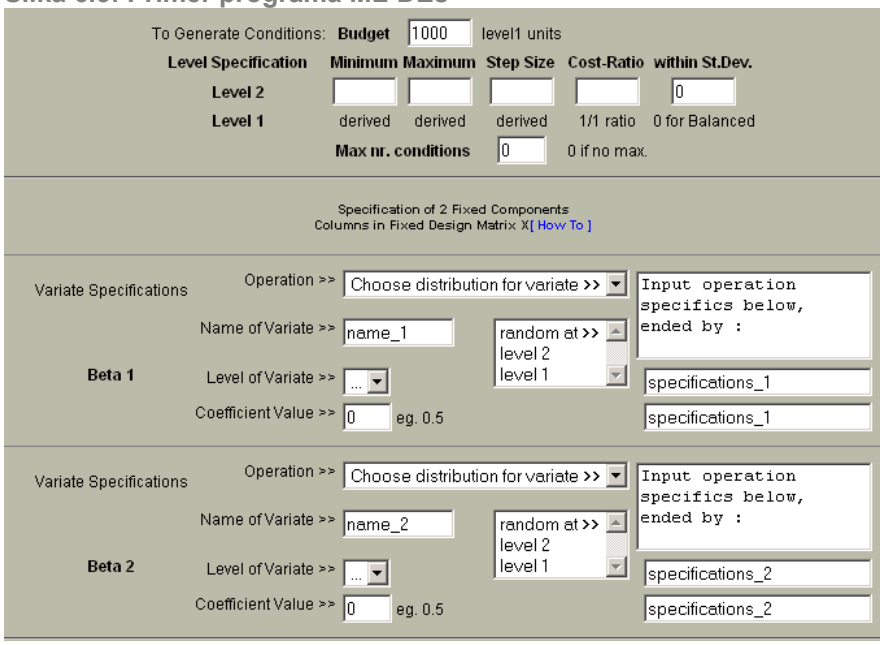
## 6.2 Programi, ki se poslužujejo simulacijske metodologije

Zaradi nezmožnosti izpeljave formul v zaprti obliki so mnogi raziskovalci uporabili simulacijske postopke, ki pa so generirani za specifičen model in za specifične vrednosti parametrov in jih tako ne moremo posploševati. V ta namen sta bili izdelani statistični orodji ML-DEs (Cools in drugi 2008) in MLPowSim (Browne in Golalizadeh 2009), ki se poslužujeta simulacij in sta precej bolj fleksibilni kot zgoraj omenjeni programski orodji.

### 6.2.1 Program ML-DEs

Statistično orodje ML-DEs[15] za raziskovanje učinkovitosti večnivojskih načrtov vključuje množico skript v programu R (R Development Core Team 2004). Te skripte omogočajo pripravo makrov za simulacijo in ocenjevanje z uporabo programa za večnivojsko modeliranje MlwiN (Rasbash in drugi 2005). Gre za dovolj enostavno in fleksibilno orodje za programiranje, ki omogoča tudi nadaljnje analize generiranih podatkov. Za dejansko uporabo programa se ne zahteva znanje R ali MlwiN, mora pa uporabnik zelo dobro razumeti parametre, ki so vključeni v analizo večnivojskih podatkov in njihovih raziskovalnih načrtov.

Slika 6.3: Primer programa ML-DEs



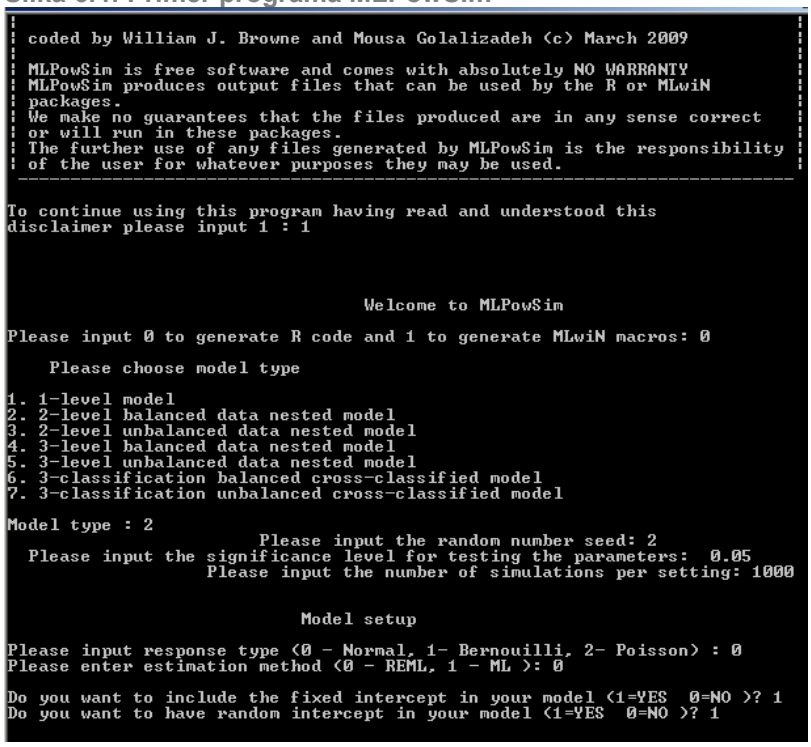

[15] Stran, namenjena ML-Des programu lahko najdemo na: http://ppw.kuleuven.be/cmes/MLDEs.html. Program je na voljo zastonj, ni pa zastonj program, v katerem se ML-Des skripte izvajajo, MlwiN. Nekatere lastnosti programa MlwiN so opisane na MlwiN strani: www.cmm.bristol.ac.uk/MLwiN/.

Program izračuna ocene parametrov in njihove standardne napake glede na dane pogoje, vključujoč osnovne statistike in podatke o konvergenci. Nadalje je izveden Waldov test, po želji tudi test deviance. Oblikovana je tudi koristna R skripta, ki reorganizira in povzema rezultate in oblikuje funkcije, ki jih lahko uporabimo za vizualizacijo podatkov. Program je zelo fleksibilen, lahko vključuje do 5 nivojev, lahko vključuje tudi neodvisne spremenljivke, ki niso normalno porazdeljene, vendar pa mora biti normalno porazdeljena vsaj odvisna spremenljivka.

### 6.2.2 Program MLPowSim

Programski paket MLPowSim ustvari R skripto in MlwiN makre, ki jih zaženemo v teh programih, in se simulirajo podatki za model kot ga je definiral uporabnik. Program lahko izvaja izračunavanje potrebne velikosti vzorca za modele z več kot dvema nivojema, za modele s prečnimi slučajnimi vplivi (*angl. crossed random effects*), neuravnotežene podatke. Za razliko od ML-DEs lahko izvedemo tudi simulacije za odvisno spremenljivko, ki ne sledi normalni porazdelitvi. Tudi za ta program je oblikovan obširen priročnik.[16]

Slika 6.4: Primer programa MLPowSim




[16] Stran, namenjena programu MLPowSim lahko najdemo na: http://seis.bris.ac.uk/~frwjb/esrc.html.
Program je na voljo zastonj, ni pa zastonj program, v katerem se skripte izvajajo, MlwiN, če se odločimo za uporabo le-tega.

# 7 Učinkovito načrtovanje velikosti vzorca na osnovi simulacij

Postopki za načrtovanje velikosti vzorca so razviti za širok nabor statističnih testov (pregled v Maxwell in drugi 2008), vendar so običajno osnovani na standardnih tehnikah, ko je zadoščeno vsem predpostavkam. Postopki za načrtovanje velikosti vzorca še niso bili razviti za nestandardne analize (npr. klasifikacija in regresijska drevesa) in/ali tehnike osnovane na računanju (*angl. computationaly based techniques*) (npr. bootstrap pristop k statističnemu sklepanju). Kljub temu, da je področje načrtovanja velikosti vzorcev za moč precej razvito, v tem času pogosto ni ustreznih metod za AIPE. Splošno načelo načrtovanja velikosti vzorca očitno drži: velikost vzorca lahko načrtujemo za katerikoli cilj ali statistično tehniko, v katerikoli situaciji z *a priori* Monte Carlo simulacijsko študijo.

*A priori* Monte Carlo simulacijska študija za načrtovanje primerne velikosti vzorca vključuje generiranje slučajnih podatkov iz populacije našega zanimanja (npr. primerne parametre, oblike porazdelitev), implementacijo določene statistične tehnike in ponavljanje večjega števila simulacij (npr. 1000) z različnimi velikostmi vzorcev. Takšen postopek nas privede do minimalne velikosti vzorca, da dosežemo nek določen cilj (npr. 90% moč, pričakovano širino intervala 0.15, 85% moč in 1% toleranco, da bo interval zaupanja dovolj ozek). Izvedba takšne *a priori* Monte Carlo simulacijske študije za načrtovanje velikosti vzorca zahteva znanje porazdelitvene oblike in populacijskih parametrov, kar pa je res tudi za tradicionalne analitične metode načrtovanja velikosti vzorcev (kjer se normalnost napak skoraj vedno predpostavlja) (Maxwell in drugi 2008).

Simulacije lahko úporabimo za aproksimacijo populacijske vzorčne porazdelitve za katerikoli parameter z empirično vzorčno porazdelitvijo. To izvedemo s ponovitvami analiziranja vsakega od dovolj velikega števila podatkov, ki smo jih simulirali pogojno glede na predpostavljene populacijske lastnosti oz. vrednosti (Muthén in Muthén 2002). Empirično vzorčno porazdelitev lahko uporabimo, da dobimo oceno natančnosti, ki jo običajno izrazimo z RMSE in združuje preciznost in pristranskost. Medtem ko pristranskost aproksimiramo z razliko med povprečjem empirične vzorčne porazdelitve in populacijsko vrednostjo, standardno napako aproksimiramo s

standardnim odklonom ocen v empirični vzorčni porazdelitvi. Empirično vzorčno porazdelitev lahko uporabimo tudi za preverjanje populacijske porazdelitve za normalnost ali katerokoli drugo referenčno porazdelitev, kot tudi za katerokoli drugo opisno statistiko kot npr. asimetričnost in sploščenost (Cools in drugi 2008).

V običajnih primerih in enostopenjskih modelih lahko moč testa natančno (ali približno) izračunamo. V nekem smislu nam moč testa pove kako pogosto bomo, glede na dane podatke, zavrnili ničelno hipotezo, ki prihaja iz določene alternative. V realnosti bomo zbrali neko množico podatkov in bomo, ali pa ne, zavrnili ničelno hipotezo. Kakorkoli, moč kot koncept prihaja iz frekventistične statistike in ima frekventistični duh v smislu, da če bi večkrat ponovili zbiranje podatkov iz iste populacije, bi lahko dobili dolgoročno povprečje pogostosti zavrnitve ničelne hipoteze; to bi ustrezalo našemu pojmu moči (Gelman in Hill 2007).

V realnosti se ne poslužujemo večkratnega zbiranja podatkov, namesto nas delo opravi računalnik, s simulacijo. Če bi lahko (večkrat) generirali podatke, ki izhajajo iz določene alternativne hipoteze, bi lahko izračunali delež zavrnjenih ničelnih hipotez oz. izračunali statistično moč. Več množic podatkov (simulacij) uporabimo, bolj natančna bo ocena. Ta pristop je še posebej privlačen, saj ponavlja postopek, ki ga bomo izvedli na dejanskih podatkih. Na ta način preverimo tudi metodo ocenjevanja, ki jo bomo uporabili in test, ki ga bomo izvedli (Gelman in Hill 2007). Za oceno statistične moči moramo poznati velikost vpliva v hipotezi. Moč nato aproksimiramo z deležem ponovitev, ki vodijo k pravilni zavrnitvi ničelne hipoteze.

Za oceno širine intervala nam ni potrebno poznati velikosti vpliva v hipotezi. S simulacijskimi postopki moramo zagotoviti le dobro oceno standardne napake in definirati še primerno širino intervala, pri čemer je zaželeno upoštevati tudi kriterij tolerance za neko določeno širino intervala. Končno nas lahko zanima tudi delež konvergence. Od vseh ponovitev simulacij izračunamo kolikšen delež je privedel do veljavne rešitve. To je odvisno od pravilnosti specifikacije modela, pa tudi od velikosti vzorca (npr. pri majhnih vzorcih je delež konvergence manjši).

Seveda se vsaka od teh aproksimacij izboljša s povečanjem števila ponovitev. Dobljena empirična vzorčna porazdelitev se s povečevanjem števila ponovitev

izboljšuje v smislu aproksimacije določenih populacijskih parametrov, vendar pa se to izboljšanje ne povečuje v neskončnost. Število vzorcev ali ponovitev moramo določiti primerno, glede na cilje analize. Glede na konkreten model mora raziskovalec raziskati stabilnost ocen standardnega odklona (npr. s tekočim povprečjem (*angl. running mean*) standardne napake za posamezen parameter) (Cools in drugi 2008).

Ko izvajamo simulacijo je zaželeno, da zagotovimo določeno množico slučajnih številk (npr. z ukazom set.seed v programu R), ki jih uporabimo za ustvarjanje vrednosti parametrov in slučajnega dela modela, kar omogoča natančno ponovitev rezultatov simulacije (npr. Cools in drugi 2008, Gelman in Hill 2007).

# 8 Rezultati simulacijske študije

Namen pričujoče magistrske naloge je preseči omejitve analitičnih metod in objavljenih simulacijskih študij s celovitejšo simulacijsko študijo za načrtovanje učinkovite večnivojske raziskave. Konkretneje bom odgovorila na vprašanje, kako je učinkovitost načrta večnivojske raziskave v smislu predvidene statistične moči in natančnosti ocenjevanja parametrov odvisna od vzorčenja na prvem in drugem nivoju, pri čemer bom preverjala tudi vpliv različnih dejavnikov, kot so različne porazdelitve varianc med nivoji (različni znotrajrazredni koeficienti), neuravnoteženi vzorčni načrti, spreminjanje velikosti vpliva. Pri vsem tem bom upoštevala tudi stroškovni vidik.

## 8.1 Simulacijski model in postopki

Uporabila sem dvonivojski model z eno pojasnjevalno spremenljivko na individualnem nivoju in eno pojasnjevalno spremenljivko na skupinskem nivoju, ki ga prikazuje enačba 3.10:

$$\gamma_{00} + \gamma_{01} z_j + \gamma_{10} x_{ij} + \gamma_{11} z_j x_{ij} + U_{0j} + U_{1j} x_{ij} + R_{ij} \qquad \textbf{(3.10)}$$

Začetni del simulacije predstavlja variiranje 3 pogojev: števila skupin, velikosti skupin ter znotrajrazrednega koeficienta.

***Število skupin*** Število skupin variira od 5 do 200 (5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 150, 200). V praksi se 50 skupin pogosto pojavlja v raziskavah organizacij ter šol, 30 pa je v splošnem sprejeto kot najmanjše možno število skupin (npr. Kreft in De Leeuw 1998).

***Velikost skupin*** Velikost skupin variira od 5 do 80 (5, 10, 20, 30, 40, 50, 80). V raziskovanju izobraževanja (*angl. educational research*) je običajna velikost skupine 30, v raziskavah družine in v longitudinalnih raziskavah pa 5.

***Znotrajrazredni koeficient*** Razpon znotrajrazrednega koeficienta (0.1, 0.2, 0.3) ustreza običajnemu rangu ICC koeficientov (Gulliford, Ukoumunne in Chinn 1999).

Skupno imamo torej 13X7X3=273 pogojev. Za vsak pogoj sem generirala 1000 simuliranih podatkovnih množic, pri čemer sem predpostavljala normalno porazdeljene ostanke. Kot njegova enonivojska različica, tudi večnivojski regresijski model predpostavlja fiksne pojasnjevalne spremenljivke. Da bi zadostili zahtevam simuliranih pogojev z najmanjšo skupno velikostjo vzorca, so množice X in Z vrednosti generirane iz standardne normalne porazdelitve. V pogoju z večjimi velikostmi vzorca, so te vrednosti ponovljene. To nam zagotavlja, da so v vseh simuliranih pogojih zvezne porazdelitve X in Z identične.

Regresijski koeficienti so določeni na sledeč način: 1.0 za presečišče, 0.3 (srednja velikost vpliva; cf. Cohen 1988) za regresijski nagib spremenljivke na mikro nivoju, 0.1 (majhna velikost vpliva) za regresijski nagib spremenljivke na makro nivoju ter 0.05 (zelo majhna velikost vpliva) za regresijski nagib člena interakcije. Varianca ostankov na mikro nivoju $\sigma^2$ je 0.5. Varianca ostankov za presečišče $\mathrm{var}(U_{0j}) = \tau_{00} = \tau_0^2$ sledi iz določenega ICC in individualne napake $\sigma^2$ (glej enačbo 3.4 za ICC). Busing (1993) sicer trdi, da sta si vpliva variance presečišča ter variance nagiba $\mathrm{var}(U_{1j}) = \tau_{11} = \tau_1^2$ podobna, vendar sem v mojem primeru izbrala manjšo varianco nagiba kot presečišča (po primeru Snijders in Bosker 1993).[17] Za poenostavitev simuliranega modela se predpostavlja, da je kovarianca med dvema u-členoma nič.

Dve funkciji največjega verjetja (ML) sta pogosti v večnivojskem ocenjevanju: FML in RML. Uporabili bomo RML, saj je le-ta skoraj vedno vsaj tako dobra kot FML, pogosto pa še boljša, posebej v ocenjevanju komponent variance (Browne 1998, Hox 2002).

[17] Znotrajrazredni koeficient 0.1 predpostavlja varianco presečišča 0.05555556 ter varianco nagiba 0.008 (varianca nagiba sorazmerno predstavlja 15% vrednost variance presečišča). Znotrajrazredni koeficient 0.2 predpostavlja varianco presečišča 0.125 ter varianco nagiba 0.019. Znotrajrazredni koeficient 0.3 predpostavlja varianco presečišča 0.2142857 ter varianco nagiba 0.032.

Podatki so simulirani in analizirani v programu R (R Development Core Team 2004). Skupno je bilo izvedenih kar 273.000 simulacij, glede na kombinacijo 7 velikosti skupin, 13 skupin ter 3 znotrajrazrednih koeficientov. Simulirani večnivojski podatki so bili analizirani s pomočjo funkcije lmer. S funkcijo lmer preverjamo prileganje linearnih hierarhičnih modelov, posplošene hierarhične linearne modele, lahko pa tudi nelinearne hierarhične modele.

V nadaljevanju bom na kratko opisala sklope po katerih bom povzela rezultate simulacijske študije.

***1.SPLOŠNE UGOTOVITVE*** Začela bom s splošnimi ugotovitvami, ki zadevajo obdelavo vseh simulacij ter vpliv števila skupin, velikosti skupin ter različnih vrednosti znotrajrazrednih koeficientov na natančnost ocen (*Poglavji 8.2 in 8.3*).

***2.MOČ STATISTIČNEGA TESTA TER ŠIRINA INTERVALOV ZAUPANJA*** V nadaljevanju se bom lotila glavnega dela; preverjanje statistične značilnosti testov ter natančnost ocene, kot jo pogojuje širina intervala zaupanja. Da zmanjšam nepotrebno kompleksnost in redundatnost, bom v nadaljevanju obravnavala samo primer znotrajrazrednega koeficienta velikosti 0.1 (*Poglavji 8.4 in 8.5*).

***3.FOKUS*** Zgornje ugotovitve bom nadgradila z vključitvijo stroškovnega vidika, s pomočjo katerega bom izluščila manjše število pogojev, ki jih bom v drugem delu bolj podrobno preverila, prikazala tudi konkretne empirične porazdelitve, itd. in se na podlagi ugotovitev odločila za najbolj optimalno (*Poglavji 8.6 in 8.7*).

***4.ANALIZA OBČUTLJIVOSTI*** V četrtem delu bom s pomočjo analize občutljivosti preverila občutljivost optimalne velikosti vzorca, kot jo določajo stroškovne omejitve, pri čemer me bo zanimala občutljivost na neuravnotežene podatke ter občutljivost na spreminjanje velikosti vpliva (*Poglavji 8.8 in 8.9*).

***PRVI DEL: splošne ugotovitve*** Pa začnimo s splošnim delom; oglejmo si splošne ugotovitve glede obnašanja simuliranih ocen glede na posamezne pogoje. Za začetek bom posebno podpoglavje namenila konvergenci, ki lahko predstavlja velik problem pri analiziranju večnivojskih podatkov, še posebej, če gre za kompleksnejše modele oz. manjše število podatkov na vseh nivojih.

## 8.2 Konvergenca in nedopustne rešitve

Od skupno 273.000 simulacij jih je konvergiralo le 84,1% (n=229.716). Postopek ocenjevanja v R-u (lmer) lahko vodi do negativnih ocen varianc. Takšne rešitve so nesprejemljive; običajen postopek je, da se takšne ocene omeji na mejno vrednost 0 (lmer takšne ocene omeji celo na izredno majhno pozitivno vrednost). Po drugi strani pa lmer privede tudi do singularne konvergence ali do napačne konvergence. Problem, ki ga lmer sicer eksplicitno ne poroča kot napako pa je popolna korelacija med slučajnimi koeficienti. Tudi te rešitve sem obravnavala kot nesprejemljive.

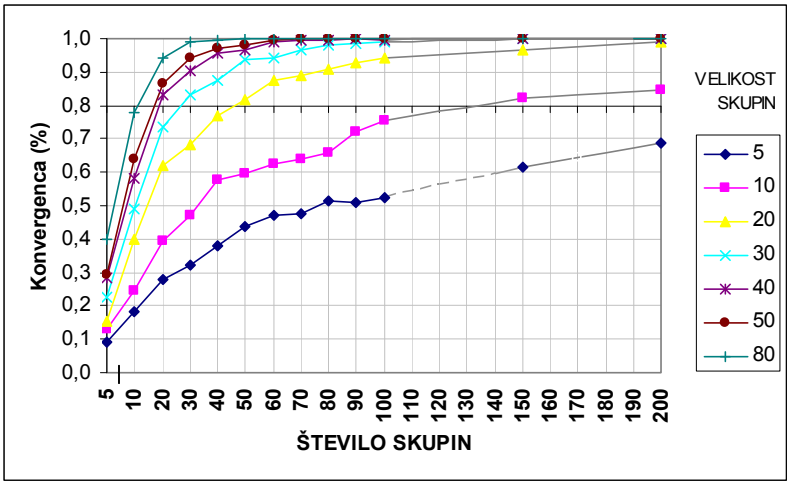
Znotrajrazredni koeficient (ICC) ima velik vpliv na konvergenco. Največ konvergiranih rešitev (90,3%) nam da znotrajrazredni koeficient 0.3 (to je seveda povezano z dejstvom, da večji znotrajrazredni koeficient pomeni večjo varianco ostankov presečišča in v našem primeru tudi nagiba).

**Tabela 8.1: Konvergenca glede na različne vrednosti ICC**

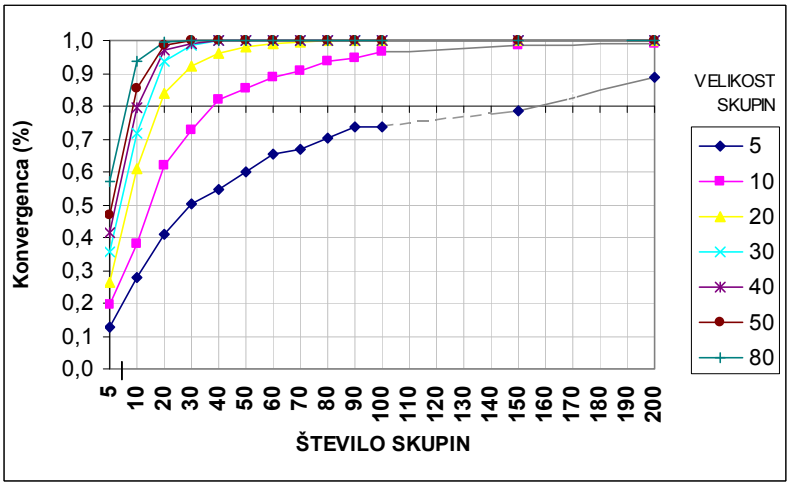
| | Število simulacij | Število konvergiranih simulacij | Delež konvergiranih simulacij |
|---|---|---|---|
| 0.1 | 91000 | 69157 | 76,0 |
| 0.2 | 91000 | 78381 | 86,1 |
| 0.3 | 91000 | 82178 | 90,3 |
| Skupaj | 273000 | 229716 | |

Velik vpliv na (ne)konvergenco ima predvsem število skupin. Za opazno izboljšanje konvergence potrebujemo vsaj 20-30 skupin, v primeru manjših velikosti skupin in/ali znotrajrazrednih koeficientov pa še več. Iz Slik 8.1a-8.1c je razvidno, da je konvergenca izraziteje slabša, ko skupine vsebujejo le 5 ali 10 enot.

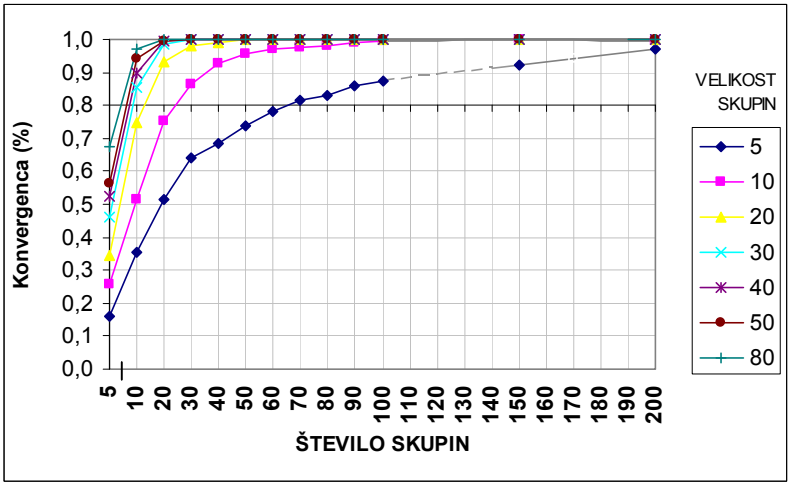
**Slika 8.1a: %konvergence glede na različne velikosti in število skupin za ICC=0.1** (vrednosti za kombinacijo števila in velikosti skupin, ki niso bile vključene v simulacijo, so v grafih interpolirane, črte, ki se nanašajo na te vrednosti pa obarvane sivo)




**Slika 8.1b: %konvergence glede na različne velikosti in število skupin za ICC=0.2**




**Slika 8.1c: %konvergence glede na različne velikosti in število skupin za ICC=0.3**




V naslednjem poglavju bom obravnavala ocene fiksnih ter slučajnih koeficientov glede na različne pogoje. Upoštevala bom seveda samo konvergirane rešitve.

## 8.3 Natančnost ocen parametrov in njihovih standardnih napak

Natančnost (*angl. accuracy*) je definirana kot ujemanje med vrednostjo, ki jo dobimo iz vzorca in dejansko, toda v glavnem neznano populacijsko vrednostjo. Formalna definicija natančnosti je dana s kvadratnim korenom povprečne kvadratne napake (RMSE), ki jo sestavljata pristranskost ter preciznost (*glej Enačba 2.2*). Preciznosti v smislu variance ne bom obravnavala, ampak jo bom obravnavala posredno v podpoglavju o standardnih napakah regresijskih koeficientov.

### 8.3.1 Regresijski koeficienti

V tem poglavju bom predstavila vpliv števila skupin, velikosti skupin ter znotrajrazrednega koeficienta na pristranskost ter preciznost v smislu standardne napake. Predstavila bom tudi natančnost standardnih napak regresijskih koeficientov.

***Pristranskost*** Pristranskost ocene je definirana kot razlika med povprečno oceno populacijskega parametra (glede na dan pogoj) ter dejansko vrednostjo populacijskega parametra ($\hat{\Theta} - \Theta$). V nadaljevanju bom primerjala povprečja s pomočjo enostranske analize variance ANOVA[18]. Rezultati le-te so predstavljeni v Tabelah 8.2a-8,2c kot p-vrednosti (Sig.) za vsak rezultat (zadnje vrstice v tabelah). Povprečna pristranskost se ne razlikuje glede na različne vrednosti ICC. Večjo pristranskost opazimo pri vseh regresijskih koeficientih v primeru velikosti skupine n=5, v primeru nagiba ter makro spremenljivke pa tudi v primeru n=10. Večjo pristranskost opazimo pri vseh regresijskih koeficientih v primeru števila skupin N=5 in N=10, v primeru presečišča pa tudi v primeru N=20 ter N=30.

[18] Uporabili smo absolutno pristranskost, saj bi v nasprotnem primeru s povprečenjem pozitivnih ter negativnih vrednosti lahko izničili dejanski vpliv.

**Primerjava povprečij absolutne pristranskosti glede na različne vrednosti znotrajrazrednega koeficienta (ICC), št. skupin (N) in velikosti skupin (n) za vse regresijske koeficiente: INT-presečišče, X-regresijski nagib spremenljivke na mikro nivoju, Z-regresijski nagib spremenljivke na makro nivoju, XZ-regresijski nagib člena interakcije)**

**Tabela 8.2a: znotrajrazredni koeficient (ICC)**

| ICC | INT | X | Z | XZ |
|---|---|---|---|---|
| 0,1 | 0,0017 | 0,0012 | 0,0023 | 0,0022 |
| 0,2 | 0,0022 | 0,0010 | 0,0032 | 0,0019 |
| 0,3 | 0,0022 | 0,0012 | 0,0033 | 0,0020 |
| Sig. | 0,451 | 0,595 | 0,337 | 0,953 |

**Tabela 8.2b: velikost skupin (n)**

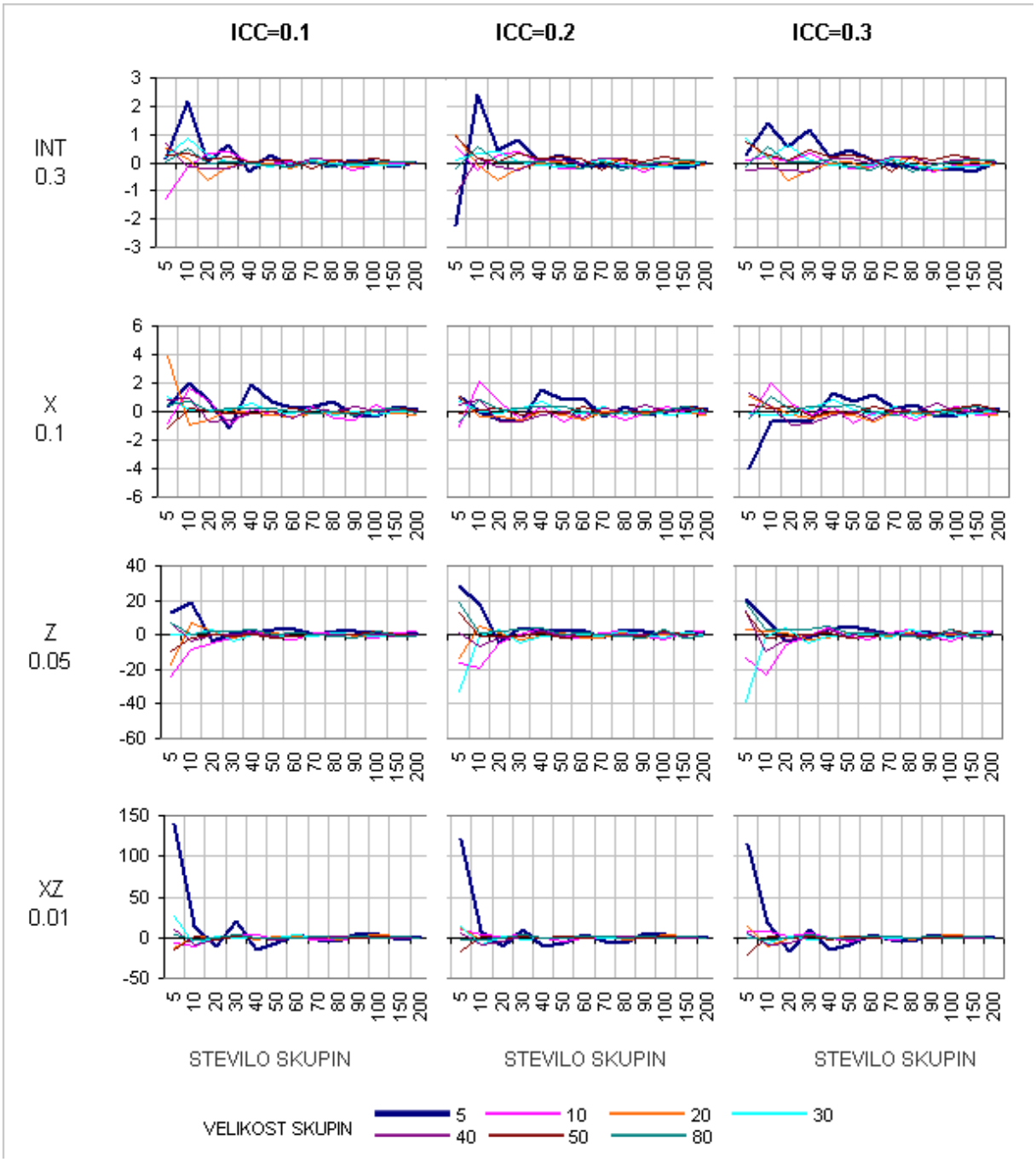
| n | INT | X | Z | XZ |
|---|---|---|---|---|
| 5 | 0,0041 | 0,0021 | 0,0045 | 0,0080 |
| 10 | 0,0018 | 0,0015 | 0,0045 | 0,0012 |
| 20 | 0,0019 | 0,0011 | 0,0022 | 0,0012 |
| 30 | 0,0017 | 0,0007 | 0,0031 | 0,0014 |
| 40 | 0,0017 | 0,0011 | 0,0018 | 0,0010 |
| 50 | 0,0019 | 0,0006 | 0,0017 | 0,0011 |
| 80 | 0,0012 | 0,0008 | 0,0027 | 0,0006 |
| Sig. | 0,001 | 0,000 | 0,064 | 0,000 |

**Tabela 8.2c: št. skupin (N)**

| N | INT | X | Z | XZ |
|---|---|---|---|---|
| 5 | 0,0058 | 0,0031 | 0,0154 | 0,0134 |
| 10 | 0,0053 | 0,0023 | 0,0068 | 0,0033 |
| 20 | 0,0027 | 0,0013 | 0,0026 | 0,0015 |
| 30 | 0,0030 | 0,0014 | 0,0019 | 0,0019 |
| 40 | 0,0008 | 0,0014 | 0,0021 | 0,0014 |
| 50 | 0,0016 | 0,0009 | 0,0013 | 0,0012 |
| 60 | 0,0012 | 0,0011 | 0,0015 | 0,0007 |
| 70 | 0,0011 | 0,0006 | 0,0008 | 0,0007 |
| 80 | 0,0013 | 0,0006 | 0,0015 | 0,0007 |
| 90 | 0,0014 | 0,0007 | 0,0010 | 0,0006 |
| 100 | 0,0010 | 0,0005 | 0,0016 | 0,0007 |
| 150 | 0,0010 | 0,0006 | 0,0007 | 0,0006 |
| 200 | 0,0004 | 0,0004 | 0,0010 | 0,0002 |
| Sig. | 0,000 | 0,000 | 0,000 | 0,000 |

V Sliki 8.2 so predstavljeni odstotki relativne pristranskosti ocene, $\frac{\hat{\Theta}-\Theta}{\Theta}\times 100$, saj nam le-ta da bolj relevantno informacijo o odklonu od prave vrednosti. Izkazalo se je, da imajo ocene fiksnih parametrov zanemarljivo pristranskost v primeru regresijskih koeficientov z večjo težo (npr. presečišče (1.0000) in nagib - mikro spremenljivka (0.30000). V tem primeru je pristranskost v relativnem smislu zanemarljiva tudi v primeru manjših vzorcev. V teh primerih je preciznost tista, ki definira natančnost. Rezultati sovpadajo s podobnimi raziskavami (npr. Maas in Hox 2005). Opazen vpliv v vseh primerih ima velikost skupin (n=5) ter število skupin (N<20). V primeru regresijskih koeficientov z manjšo težo (makro spremenljivka (0.10) ter interakcijske spremenljivke (0.05)) pa je opazna znatna relativna pristranskost v primeru manjšega števila skupin (velikost skupine manjša od 20) ter manjših velikosti skupin.

**Slika 8.2: Relativna pristranskost (%) regresijskih koeficientov glede na število ter velikost skupin**




***Preciznost: standardne napake, variabilnost koeficientov*** Preciznost regresijskih koeficientov lahko izračunamo na dva načina; lahko jo izrazimo s standardnim odklonom, ki ga izračunamo iz empirične vzorčne porazdelitve glede na posamezen pogoj (ICC, število skupin, velikost skupin) ali pa jo izrazimo s kvadriranjem standardne napake, kot jo izračuna program. V tem poglavju bomo na preciznost gledali z vidika standardnega odklona. Standardna napaka kot jo izračuna lmer je

izračunana iz rezidualov, pri predpostavki, da je model ustrezno specificiran. Takšne standardne napake odražajo negotovost in tako tudi varianco, vkolikor bi ponovili raziskavo z drugim vzorcem iz iste populacije. Standardni odklon ocen je samo drug način ocenjevanja te negotovosti glede rezultatov ponovitve iste raziskave z drugim vzorcem. Z drugimi besedami, gre za oceni iste stvari in s pravilnim modelom bi morali, če bi uporabili različna vzorca, obe predstavljati zelo podobno vrednost. Tudi v našem primeru gre za podobni oceni, predvsem, ko govorimo o večjem številu skupin (20 ali več)[19].

V tabelah 8.3a-8.3c je prikazana primerjava povprečij standardnih napak po posameznih pogojih s pomočjo enostranske analize variance ANOVA. Rezultati so predstavljeni kot p-vrednosti (Sig.) za vse primerjave povprečij (zadnje vrstice v tabelah). Standardna napaka se povečuje z večanjem znotrajrazrednega korelacijskega koeficienta ter z zmanjševanjem števila skupin ter velikosti skupin. Standardne napake regresijskih koeficientov so znatno višje v primeru, ko je skupin manj kot 20, nekoliko manjši vpliv ima velikost skupin, predvsem ko imamo v skupini le 5 enot. Večji vpliv ICC lahko opazimo v primeru presečišča ter spremenljivke na makro nivoju, večji vpliv velikosti skupin pa v primeru nagiba sprmenljivke na mikro nivoju ter člena interakcije.

**Primerjava povprečij standardnih napak glede na različne vrednosti znotrajrazrednega koeficienta (ICC), št. skupin (N) in velikosti skupin (n)**

**Tabela 8.3a: št. skupin (N)**

| N | INT | X | Z | XZ |
|---|---|---|---|---|
| 5 | 0,212 | 0,127 | 0,252 | 0,153 |
| 10 | 0,132 | 0,074 | 0,143 | 0,080 |
| 20 | 0,089 | 0,049 | 0,093 | 0,051 |
| 30 | 0,073 | 0,040 | 0,074 | 0,041 |
| 40 | 0,062 | 0,034 | 0,064 | 0,035 |
| 50 | 0,056 | 0,031 | 0,057 | 0,031 |
| 60 | 0,051 | 0,028 | 0,052 | 0,029 |
| 70 | 0,047 | 0,026 | 0,048 | 0,026 |
| 80 | 0,044 | 0,024 | 0,045 | 0,025 |
| 90 | 0,042 | 0,023 | 0,042 | 0,023 |
| 100 | 0,039 | 0,022 | 0,040 | 0,022 |
| 150 | 0,032 | 0,018 | 0,032 | 0,018 |
| 200 | 0,028 | 0,016 | 0,028 | 0,016 |
| Sig. | 0,000 | 0,000 | 0,000 | 0,000 |

**Tabela 8.3b: velikost skupin (n)**

| n | INT | X | Z | XZ |
|---|---|---|---|---|
| 5 | 0,069 | 0,057 | 0,072 | 0,059 |
| 10 | 0,062 | 0,042 | 0,065 | 0,044 |
| 20 | 0,059 | 0,033 | 0,061 | 0,035 |
| 30 | 0,058 | 0,030 | 0,061 | 0,032 |
| 40 | 0,058 | 0,029 | 0,061 | 0,030 |
| 50 | 0,058 | 0,028 | 0,061 | 0,029 |
| 80 | 0,059 | 0,026 | 0,061 | 0,027 |
| Sig. | 0,000 | 0,000 | 0,000 | 0,000 |

**Tabela 8.3c: znotrajrazredni koeficient (ICC)**

| ICC | INT | X | Z | XZ |
|---|---|---|---|---|
| 0,1 | 0,041 | 0,027 | 0,043 | 0,028 |
| 0,2 | 0,059 | 0,033 | 0,062 | 0,035 |
| 0,3 | 0,077 | 0,039 | 0,080 | 0,041 |
| Sig. | 0,000 | 0,000 | 0,000 | 0,000 |

INT-presečišče, X-regresijski nagib spremenljivke na mikro nivoju, Z-regresijski nagib spremenljivke na makro nivoju, XZ-regresijski nagib člena interakcije)

[19] V ta namen sem preverila povezanost spremenljivk glede na dva načina. Korelacijski koeficienti so pokazali na močno povezanost (r>0.95 (Sig.<0.01)). Pri 20 skupinah je povezanost skoraj popolna.

***Natančnost standardnih napak regresijskih koeficientov*** Preciznost lahko uporabimo za definiranje intervala zaupanja za oceno. Predpostavka o normalni porazdelitvi in nominalni Napaki tipa I z verjetnostjo $\alpha$ nam omogoča definiranje $(1-\alpha)*100\%$ intervalov zaupanja, za katere pričakujemo, da bodo vključevali vrednost parametra z verjetnostjo $(1-\alpha)$. Za oceno natančnosti standardnih napak je bil za vsak parameter v vsaki simulirani podatkovni množici ustvarjen 95% interval zaupanja z uporabo *asimptotične standardne normalne porazdelitve* (cf. Goldstein 1995). Za vsak parameter sem oblikovala dihotomno spremenljivko za nepokritost, ki je bila enaka nič, če je prava vrednost ležala v intervalu zaupanja in enaka ena, če je bila prava vrednost izven intervala zaupanja.

Vpliv števila skupin na nepokritost je predstavljena v Tabeli 8.4a, vpliv velikosti skupin na nepokritost je predstavljen v Tabeli 8.4b, vpliv ICC na nepokritost pa je predstavljen v Tabeli 8.4c. Za preverjanje vpliva različnih simulacijskih pogojev na nepokritost sem uporabila metodo ANOVA.[20] Rezultati so predstavljeni v Tabelah 8.4a-8.4c kot p-vrednosti (Sig.) za vsak rezultat (zadnje vrstice v tabelah).

**Tabela 8.4a: Nepokritost 95% intervala zaupanja glede na število skupin (N)**

| N | INT | X | Z | XZ |
|---|---|---|---|---|
| 5 | 0,116 | 0,056 | 0,107 | 0,052 |
| 10 | 0,081 | 0,058 | 0,075 | 0,055 |
| 20 | 0,066 | 0,054 | 0,065 | 0,055 |
| 30 | 0,063 | 0,054 | 0,056 | 0,053 |
| 40 | 0,056 | 0,054 | 0,057 | 0,055 |
| 50 | 0,057 | 0,051 | 0,054 | 0,053 |
| 60 | 0,055 | 0,047 | 0,054 | 0,051 |
| 70 | 0,056 | 0,055 | 0,057 | 0,057 |
| 80 | 0,049 | 0,052 | 0,055 | 0,053 |
| 90 | 0,054 | 0,054 | 0,050 | 0,049 |
| 100 | 0,053 | 0,053 | 0,049 | 0,053 |
| 150 | 0,050 | 0,047 | 0,054 | 0,050 |
| 200 | 0,049 | 0,050 | 0,052 | 0,053 |
| Sig. | 0,000 | 0,000 | 0,000 | 0,013 |

INT-presečišče, X-regresijski nagib spremenljivke na mikro nivoju, Z-regresijski nagib spremenljivke na makro nivoju, XZ-regresijski nagib člena interakcije)

**Tabela 8.4b: Nepokritost 95% intervala zaupanja glede na velikost skupin (n)**

| n | INT | X | Z | XZ |
|---|---|---|---|---|
| 5 | 0,061 | 0,045 | 0,057 | 0,045 |
| 10 | 0,060 | 0,051 | 0,055 | 0,048 |
| 20 | 0,056 | 0,051 | 0,055 | 0,052 |
| 30 | 0,058 | 0,052 | 0,057 | 0,054 |
| 40 | 0,055 | 0,052 | 0,058 | 0,054 |
| 50 | 0,060 | 0,054 | 0,058 | 0,055 |
| 80 | 0,061 | 0,058 | 0,060 | 0,059 |
| Sig. | 0,001 | 0,000 | 0,126 | 0,000 |

**Tabela 8.4c: Nepokritost 95% intervala zaupanja glede na znotrajrazredni koeficient (ICC)**

| ICC | INT | X | Z | XZ |
|---|---|---|---|---|
| 0,1 | 0,056 | 0,049 | 0,056 | 0,050 |
| 0,2 | 0,059 | 0,053 | 0,057 | 0,053 |
| 0,3 | 0,060 | 0,055 | 0,058 | 0,056 |
| Sig. | 0,008 | 0,000 | 0,134 | 0,000 |

[20]Glede na to, da imamo 229.716 simuliranih pogojev, imamo ogromno moč testa. Glede na to, da imamo cca. 1000 simulacij na vsak pogoj, je standardna napaka za verjetnosti nepokritosti približno 0.007. Kot rezultat, na standardni stopnji značilnosti α=0.05, postanejo statistično značilne že ekstremno majhne napake nepokritosti. Zato sem se odločila, da v tem primeru za glavne vplive simuliranih pogojev spremenim kriterij za značilnost v α=0.01.

Nominalna stopnja nepokritosti je 5%. Lahko vidimo, da je vpliv števila skupin na standardne napake regresijskih koeficientov majhen, ko gre za število skupin 20 ali več (presečišče in regresijski nagib spremenljivke na makro nivoju). V primeru manjšega števila skupin je v primeru teh dveh spremenljivk nepokritost večja od 7.5%. Kljub temu, da nepokritost ni zelo različna od nominalne, je 95% interval zaupanja v teh dveh primerih preozek (standardne napake so ocenjene kot premajhne, saj izračunani intervali zaupanja v premajhni meri vključujejo pravo vrednost). V primeru drugih dveh regresijskih koeficientov je sicer opaziti odklon, vendar v precej manjši meri (večja standardna napaka, širši interval zaupanja). Vpliv velikosti skupin na nepokritost je zelo majhen, celo zanemarljiv. Vpliv velikosti znotrajrazrednega koeficienta na standardne napake regresijskih koeficientov je jasen, a zanemarljiv. V obeh zadnjih primerih razlike niso statistično značilne v primeru regresijskega nagiba spremenljivke na makro nivoju (Z).

### 8.3.2 Slučajni parametri

Poglejmo še kako izbrani pogoji vplivajo na pristranskost ter preciznost slučajnih parametrov.

***Pristranskost*** Za ocene varianc obstaja večja pristranskost kot za regresijske koeficiente. Predpostavka o normalnosti je v teh primerih običajno nerealistična. Slučajni parametri kažejo večjo občutljivost na število skupin ter velikost skupin kot regresijski parametri. V tabelah 8.5a-8.5c je prikazana primerjava povprečij po posameznih pogojih s pomočjo enostranske analize variance ANOVA[21]. Kot pri fiksnih koeficientih se povprečna absolutna pristranskost ne razlikuje glede na različne vrednosti ICC. Znatno pristranskost opazimo pri vseh slučajnih koeficientih v primeru velikosti skupine n=5, nekoliko povečano tudi v primeru n=10 (razen pri kovarianci). Nadpovprečno pristranskost opazimo pri vseh slučajnih koeficientih predvsem v primeru števila skupin N=5, pa tudi N=10.

[21] Uporabili smo absolutno pristranskost, saj bi v nasprotnem primeru s povprečenjem pozitivnih ter negativnih vrednosti lahko izničili dejanski vpliv.

**Primerjava povprečij absolutne pristranskosti glede na različne vrednosti znotrajrazrednega koeficienta (ICC), št. skupin (N) in velikosti skupin (n)**

**Tabela 8.5a: znotrajrazredni koeficient (ICC)**

| ICC | u0 | u1 | rij | cov |
|---|---|---|---|---|
| 0,1 | 0,00441 | 0,00851 | 0,00697 | 0,00032 |
| 0,2 | 0,00440 | 0,00713 | 0,00483 | 0,00057 |
| 0,3 | 0,00512 | 0,00677 | 0,00376 | 0,00057 |
| Sig. | 0,929 | 0,869 | 0,263 | 0,161 |

**Tabela 8.5b: velikost skupin (n)**

| n | u0 | u1 | rij | cov |
|---|---|---|---|---|
| 5 | 0,01279 | 0,03266 | 0,02323 | 0,00102 |
| 10 | 0,00803 | 0,01034 | 0,00771 | 0,00044 |
| 20 | 0,00422 | 0,00377 | 0,00269 | 0,00061 |
| 30 | 0,00270 | 0,00225 | 0,00113 | 0,00040 |
| 40 | 0,00182 | 0,00157 | 0,00076 | 0,00049 |
| 50 | 0,00137 | 0,00109 | 0,00040 | 0,00029 |
| 80 | 0,00159 | 0,00062 | 0,00041 | 0,00015 |
| Sig. | 0,001 | 0,000 | 0,000 | 0,007 |

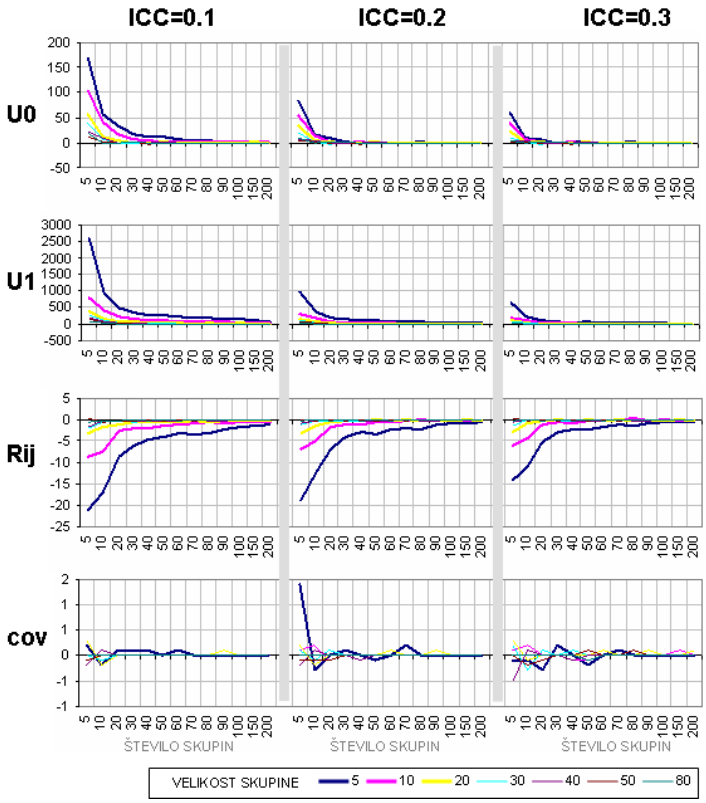
**Tabela 8.5c: število skupin (N)**

| N | u0 | u1 | rij | cov |
|---|---|---|---|---|
| 5 | 0,03780 | 0,04789 | 0,02247 | 0,00197 |
| 10 | 0,00872 | 0,01821 | 0,01517 | 0,00121 |
| 20 | 0,00404 | 0,00834 | 0,00710 | 0,00051 |
| 30 | 0,00185 | 0,00497 | 0,00442 | 0,00038 |
| 40 | 0,00168 | 0,00370 | 0,00374 | 0,00030 |
| 50 | 0,00168 | 0,00340 | 0,00337 | 0,00042 |
| 60 | 0,00068 | 0,00265 | 0,00249 | 0,00031 |
| 70 | 0,00079 | 0,00201 | 0,00215 | 0,00025 |
| 80 | 0,00071 | 0,00181 | 0,00218 | 0,00017 |
| 90 | 0,00085 | 0,00150 | 0,00153 | 0,00030 |
| 100 | 0,00065 | 0,00123 | 0,00111 | 0,00017 |
| 150 | 0,00056 | 0,00082 | 0,00084 | 0,00017 |
| 200 | 0,00037 | 0,00061 | 0,00088 | 0,00014 |
| Sig. | 0,000 | 0,000 | 0,000 | 0,000 |

*u0-varianca presečišča, u1-varianca nagiba, rij-individualna napaka, cov-kovarianca med varianco presečišča in varianco nagiba

V spodnji sliki so predstavljeni odstotki relativne pristranskosti ocene. Opazimo veliko občutljivost individualne napake (Rij) na velikost skupine (še posebej, ko je n<20). V tem primeru je opazna sistematična pristranskost navzdol. Kovarianca ima majhno relativno pristranskost, nekoliko izstopa velikost skupine n=5.

**Slika 8.6: Relativna pristranskost slučajnih koeficientov glede na število ter velikost skupin**



***Preciznost: standardne napake, variabilnost koeficientov*** Funkcija lmer v programu R sicer ne vrne standardnih napak za slučajne koeficiente, jih pa lahko v okviru simulacijskih ponovitev izračunamo iz empiričnih vzorčnih porazdelitev (standardni odklon porazdelitve). Tako nimamo standardnih napak za vsako posamezno simulacijo, ampak samo za vsak posamezen pogoj. V Tabelah 8.6a-8.6c je prikazana primerjava povprečij po posameznih pogojih s pomočjo enostranske analize variance ANOVA[22]. Standardna napaka se v povprečju povečuje z večanjem ICC (statistično značilno v primeru variance presečišča ter individualne variance) ter z zmanjšanjem števila skupin ter velikosti skupin. Lahko vidimo, da izstopa predvsem število skupin, N<20 ter velikost skupin, n<20.

**Primerjava povprečij standardnih napak vrednosti znotrajrazrednega koeficienta (ICC), št. skupin (N) in velikosti skupin (n)**

**Tabela 8.6a: znotrajrazredni koeficienta (ICC)**

| ICC | u0 | u1 | rij | cov |
|---|---|---|---|---|
| 0,1 | 0,05308 | 0,03857 | 0,05696 | 0,04184 |
| 0,2 | 0,07037 | 0,04305 | 0,07536 | 0,04657 |
| 0,3 | 0,08754 | 0,04788 | 0,09351 | 0,05164 |
| Sig. | 0,000 | 0,342 | 0,001 | 0,455 |

**Tabela 8.6b: velikost skupin (n)**

| n | u0 | u1 | rij | cov |
|---|---|---|---|---|
| 5 | 0,08859 | 0,07867 | 0,09681 | 0,08738 |
| 10 | 0,07666 | 0,05416 | 0,08195 | 0,05812 |
| 20 | 0,06985 | 0,04115 | 0,07427 | 0,04393 |
| 30 | 0,06600 | 0,03572 | 0,07042 | 0,03845 |
| 40 | 0,06455 | 0,03294 | 0,06915 | 0,03554 |
| 50 | 0,06373 | 0,03113 | 0,06759 | 0,03321 |
| 80 | 0,06290 | 0,02840 | 0,06675 | 0,03017 |
| Sig. | 0,328 | 0,000 | 0,385 | 0,000 |

**Tabela 8.6c: število skupin (n)**

| N | u0 | u1 | rij | cov |
|---|---|---|---|---|
| 5 | 0,21686 | 0,15022 | 0,25913 | 0,18169 |
| 10 | 0,13173 | 0,08311 | 0,14338 | 0,09079 |
| 20 | 0,09004 | 0,05387 | 0,09379 | 0,05627 |
| 30 | 0,07280 | 0,04255 | 0,07467 | 0,04376 |
| 40 | 0,06270 | 0,03648 | 0,06388 | 0,03724 |
| 50 | 0,05617 | 0,03249 | 0,05712 | 0,03304 |
| 60 | 0,05112 | 0,02946 | 0,05186 | 0,02993 |
| 70 | 0,04728 | 0,02709 | 0,04781 | 0,02742 |
| 80 | 0,04421 | 0,02533 | 0,04467 | 0,02562 |
| 90 | 0,04166 | 0,02380 | 0,04202 | 0,02402 |
| 100 | 0,03950 | 0,02257 | 0,03983 | 0,02276 |
| 150 | 0,03225 | 0,01835 | 0,03240 | 0,01845 |
| 200 | 0,02793 | 0,01585 | 0,02804 | 0,01591 |
| Sig. | 0,000 | 0,000 | 0,000 | 0,000 |

*u0-varianca presečišča, u1-varianca nagiba, rij-individualna napaka, cov-kovarianca med varianco presečišča in varianco nagiba

***Natančnost standardnih napak slučajnih koeficientov*** Natančnosti standardnih napak slučajnih koeficientov v pričujoči nalogi nisem ugotavljala, so pa raziskave (npr. Maas in Hox 2005) z uporabo standardnih napak kot jih vrnejo različni programi (npr. SAS, MlWin) pokazale, da je vpliv števila skupin na standardne napake komponent variance nedvomno večji kot pri fiksnih koeficientih. Čeprav nepokritost v teh primerih ni zelo napačna, je v primeru slučajnih parametrov 95% interval zaupanja očitno preozek.

[22] Uporabili smo absolutno pristranskost, saj bi v nasprotnem primeru s povprečenjem pozitivnih ter negativnih vrednosti lahko izničili dejanski vpliv.

Velikost nepokritosti v teh primerih implicira, da so standardne napake za variance na makro nivoju ocenjene približno 15% premajhne s 30 skupinami, s 50 skupinami pa približno 9% premajhne. Tudi v primeru velikosti skupin gre za podobno ugotovitev; pokrivanje 95% intervalov zaupanja je dovolj dobro za regresijske koeficiente, slabše za variance. Stopnje pokritosti se izboljšajo s povečevanjem velikosti skupin, vendar pa ima velikost skupin manjši vpliv na stopnjo pokritosti kot število skupin. Nepokritost varianc na makro nivoju se ne izboljša s povečanjem velikosti skupin. Razlika v znotrajrazrednem koeficientu nima vpliva na stopnjo pokritosti. V primeru zelo majhnih vzorcev (10 skupin velikosti 5) so se standardne napake izkazale za premajhne; stopnja nepokritosti za variance na makro nivoju so rangirale od 16.3% do 30.4%. Standardne napake varianc na makro nivoju so se pri tako majhnih vzorcih izkazale za nesprejemljive.

***DRUGI DEL: Preverjanje statistične značilnosti testov in natančnosti ocen***

Preverila in predstavila sem osnovne rezultate, ki lahko močno olajšajo interpretacijo nadaljnjih rezultatov. V nadaljevanju se bom lotila glavnega dela; preverjanje statistične značilnosti testov ter natančnost ocen, kot jo pogojuje širina intervala. Da zmanjšam nepotrebno kompleksnost in redundatnost bom v nadaljevanju obravnavala samo primer znotrajrazrednega koeficienta velikosti 0.1[23].

## 8.4 Testiranje statistične značilnosti / Moč testa

Ko želimo odgovoriti na določeno raziskovalno vprašanje pričakujemo, da je ničelna hipoteza napačna in da jo bomo na osnovi naših podatkov lahko zavrnili. Če imamo, glede na našo domnevno pravo oceno test hipoteze z nizko močjo, kljub pravilnosti naše alternativne hipoteze pogosto ne bomo mogli zavrniti ničelne hipoteze. Za splošno uporabo se priporoča specifikacija moči .80 (β=.20) oz. .90 (β=.10).

V pričujoči nalogi sem računala moč s pomočjo dveh metod, ki sta se izkazali za zelo podobni. Gre za t.i. metodo nič/ena ter metodo standardne napake.

***Metoda Nič/Ena*** Metoda je podobna Waldovemu testu. V vsaki simulaciji dobimo oceno za vsak parameter, ki nas zanima in ustrezno standardno napako. Nato lahko izračunamo (Gaussov) interval zaupanja za parameter. Če interval ne vsebuje ničle, lahko zavrnemo ničelno hipotezo in damo tej simulaciji točko 1. Če pa interval vsebuje 0, ničelne hipoteze ne moremo zavrniti in simulacija dobi točko 0. Moč predstavlja povprečje točk glede na ustrezno kombinacijo pogojev (Browne in drugi 2009).

***Metoda standardne napake*** Slaba stran prve metode je, da potrebujemo veliko simulacij, če želimo dobiti natančno oceno moči. Alternativna metoda (priporoča jo Joop Hox 2002) je, da pogledamo standardno napako za vsako simulacijo. Če vzamemo povprečje teh standardnih napak po

[23]Kot sem že omenila sem se za simulacijski model odločila glede na primer iz članka Snijders in Bosker (1993), kjer je uporabljen takšno razmerje variance presečiščča ter nagiba.

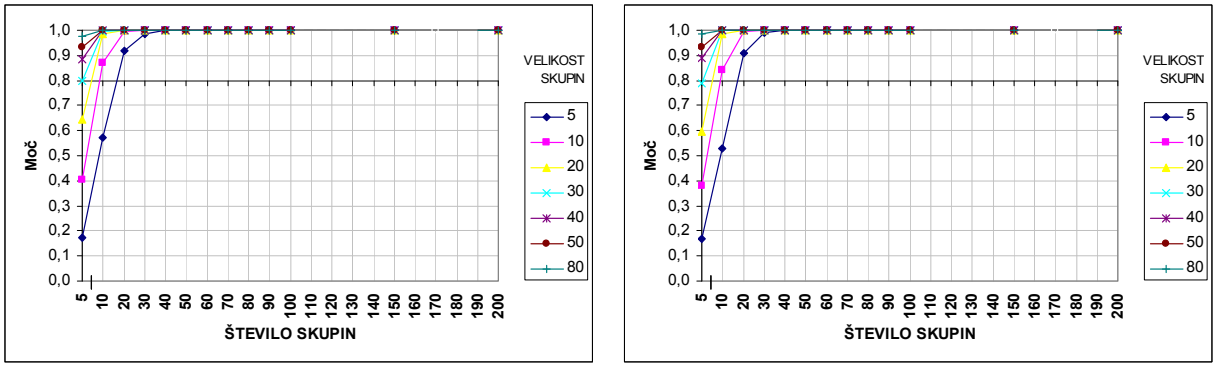
različnih množicah simulacij, skupaj s pravo vrednostjo parametra in stopnjo značilnosti α, lahko uporabimo Enačbo 2.1. in rešimo enačbo za moč (1-β).

### 8.4.1 Regresijski koeficienti

V naslednjih grafih je predstavljena moč testa za regresijske koeficiente glede na dve metodi. Odločila sem se, da kot zadostno moč uporabim moč 0.8.[24] Vrednosti za kombinacijo števila in velikosti skupin, ki niso bile vključene v simulacijo, so v grafih interpolirane, črte, ki se nanašajo na te vrednosti, pa so obarvane sivo.

**Slika 8.7: Moč testa za regresijski koeficient spremenljivke na mikro nivoju po metodi Nič/Ena (0/1) ter po metodi standardne napake (SE)**

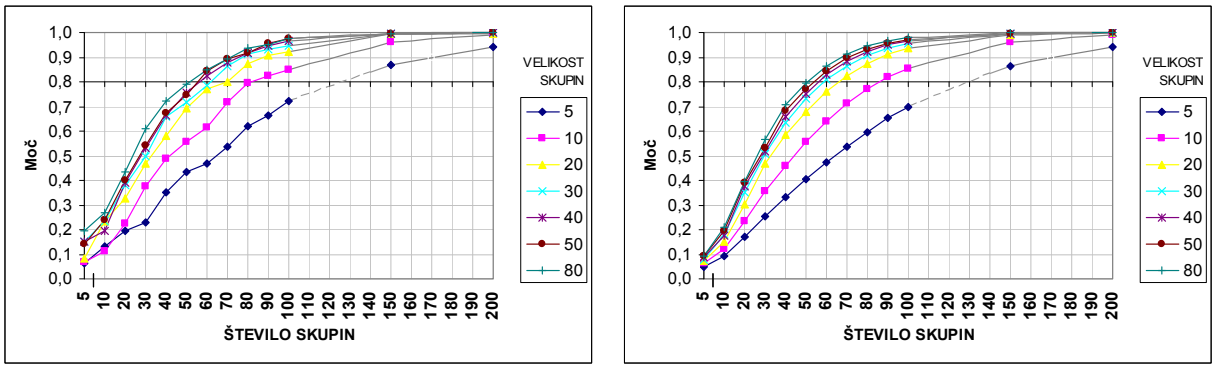
**0/1** **SE**




V primeru spremenljivke na mikro nivoju (njena populacijska vrednost je 0.3) (*glej Slika 8.7*) zadostno moč dosežemo že pri 5 skupinah z velikostjo skupin 30 oz. pri 10 skupinah z velikostjo skupin 10. Pri 20 skupinah je dovolj že 5 skupin. Razlike med dvema metodama so neopazne.

**Slika 8.8: Moč testa za regresijski koeficient spremenljivke na makro nivoju po metodi Nič/Ena (0/1) ter po metodi standardne napake (SE)**

**0/1** **SE**

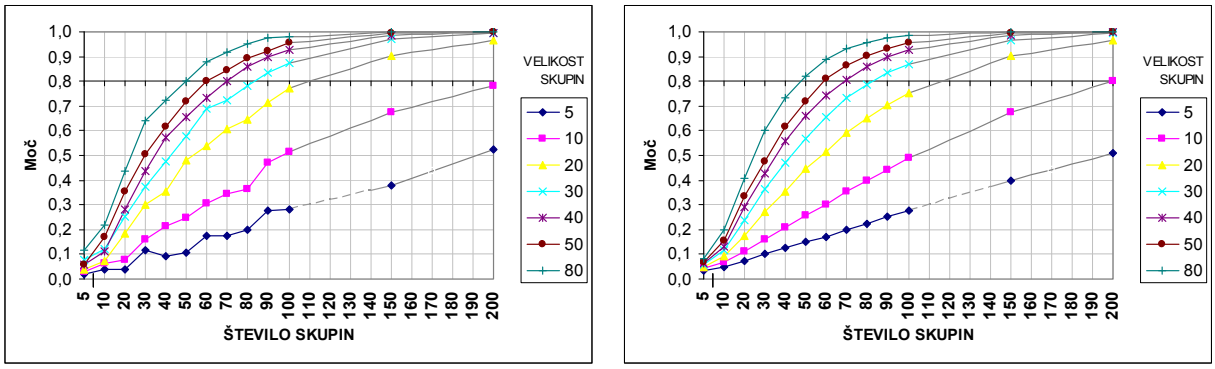



[24] Ker statistično testiranje značilnosti presečišča običajno ni primarni cilj raziskovalca, rezultatov za ta parameter na tem mestu ne podajam.

V primeru spremenljivke na makro nivoju (njena populacijska vrednost je 0.1) (*glej Slika 8.8*) dosežemo zadostno moč pri 50 skupinah z velikostjo skupin 80, pri 60 skupinah z velikostjo skupin 30, pri 70 skupinah z velikostjo 20, pri 80 skupinah z velikostjo 10, pri cca. 130 skupinah z velikostjo 5. Lahko vidimo, da so zaporedne vrednosti, ki jih dobimo z metodo 0/1 precej bolj variabilne (predvsem ko gre za manjše velikosti skupin (n<30)), kot z metodo standardne napake, kjer dobimo veliko bolj zglajene linije.

**Slika 8.9: Moč testa za regresijski koeficient interakcijske spremenljivke po metodi Nič/Ena (0/1) ter po metodi standardne napake (SE)**

**0/1** **SE**



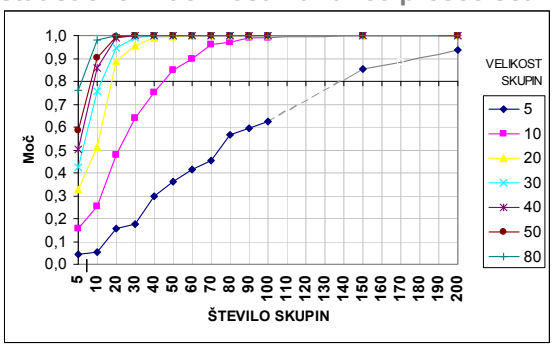

V primeru člena interakcije (prava vrednost je 0.05) (*glej Slika 8.9*) dosežemo zadostno moč pri 50 skupinah z velikostjo skupin 80, pri 60 skupinah z velikostjo skupin 50, pri 70 skupinah z velikostjo 40, pri 80 skupinah z velikostjo 30, pri cca 120 skupinah z velikostjo 20, pri 200 skupinah z velikostjo 5. Tudi v tem primeru dobimo z metodo standardne napake veliko bolj zglajene linije, predvsem ko gre za manjše velikosti vzorcev, zaporedne vrednosti višjih skupin pa so lahko celo nižje kot predhodne (*glej razmerje pri velikosti skupine 5 od 30 na 40 skupin*).

### 8.4.2 Slučajni parametri

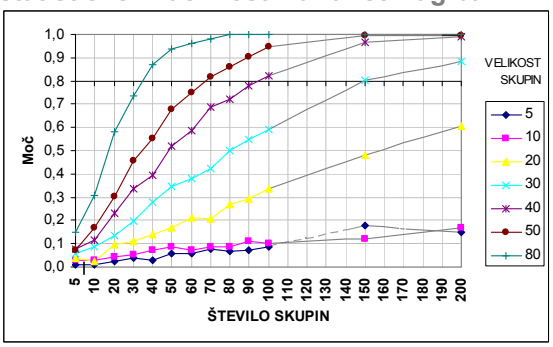
Za testiranje variančnih parametrov se ponavadi ne priporoča Waldov test, saj lahko uporablja podcenjeno standardno napako. Namesto Waldovega testa lahko uporabimo Test deviance oz. Test razmerja verjetij (*angl. likelihood ratio test (LR)*), za katerega testno statistiko interpretiramo glede na kombinacijo $\chi^2$ porazdelitev s prostostnimi stopnjami, ki so odvisne od števila variančnih parametrov in od uporabljenega algoritma ocenjevanja (Stram in Lee 2000). LR test je bil izveden za statistično testiranje variance nagiba. Testna statistika se interpretira z uporabo enakih kombinacij $\chi^2$ porazdelitev z dvema in nič prostostnimi stopnjami, ker sta na drugem nivoju slučajna samo nagib ter presečišče. Ko varianco izločimo iz modela, izločimo tudi vse z njo povezane kovariance. Za testiranje same kovariance (ki je tu nismo izvajali) bi uporabili $\chi^2$ porazdelitev z eno samo prostostno stopnjo.

V Sliki 8.10 so prikazani deleži zavrnitev ničelne hipoteze (Moč) za varianco presečišča. Če privzamemo enak kriterij kot pri regresijskih koeficientih (to je 0.8), lahko obravnavamo kot zadosten vzorec za testiranje variance presečišča že precej majhno število skupin, če je le velikost skupin dovolj velika.

**Slika 8.10: Test deviance za testiranje statistične značilnosti variance presečišča**




**Slika 8.11: Test deviance za testiranje statistične značilnosti variance nagiba**




Za testiranje statistične značilnosti variance nagiba (*glej Slika 8.11*) potrebujemo precej večje število skupin kot za testiranje statistične značilnosti variance presečišča. Šele s 30 skupinami (kar je v splošnem priporočeno kot minimalno število skupin) dosežemo 80% verjetnost za statistično značilnost variance nagiba, pa še to z velikostjo skupin večjo od 60.

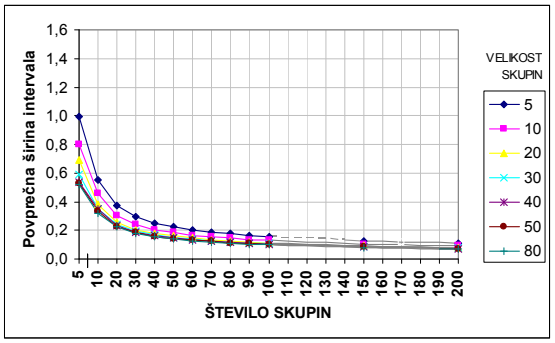
## 8.5 Intervali zaupanja

Namesto da raziskovalec preprosto testira ali je dana ocena parametra neka točna in vnaprej določena vrednost, da oblikovanje 100(1-α)% intervala zaupanja za preučevani parameter pogosto bolj smiselno informacijo. Intervali zaupanja lahko zagotovijo raziskovalcu visoko stopnjo gotovosti, da leži prava vrednost parametra znotraj nekih meja zaupanja. Razumevanje verjetnega obsega vrednosti parametra vodi do boljšega razumevanja preučevanega pojava kot samo preprosto sklepanje ali je parameter statistično značilen ali ne. Glede na perspektivo natančnosti ocen parametrov nas čim ožji interval zaupanja vodi v večjo gotovost, da opazovana ocena parametra v večji meri aproksimira ustrezen populacijski parameter (Kelley in Maxwell 2003). Denimo, da želimo biti 80% prepričani, da dobimo interval, ki ni širši od 0.15, razen v primeru interakcijske spremenljivke, kjer bi želeli, da je širina intervala največ 0.09. V tem primeru bo toleranca enaka 0.20. Takšen cilj jasno zahteva večjo velikost vzorca kot če smo pripravljeni tolerirati samo pričakovano (povprečno) širino intervala, ki naj bi bila dovolj ozka.

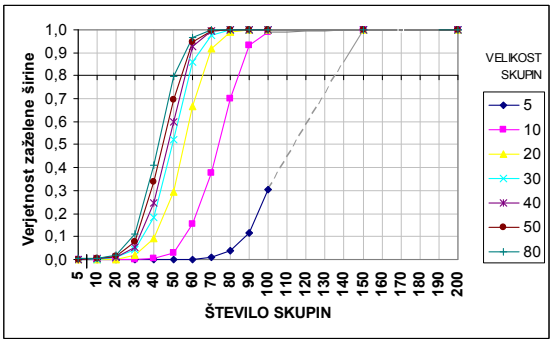
### 8.5.1 Regresijski koeficienti

Povprečne širine intervalov za presečišče so prikazane na Sliki 8.12, na Sliki 8.13 pa so prikazani deleži širine intervala, ki ne presegajo vrednosti 0.15 znotraj posameznih 1000 simulacij na pogoj, po številu skupin in velikostih skupine. Želimo biti 80% prepričani, da dobimo interval, ki ni širši od 0.15. To se zgodi šele v primeru 50 skupin, pa še to, če je v vsaki skupini kar 80 enot. S 60 skupinami bi potrebovali vsaj 30 enot v skupini, s 70 skupinami 20 enot, itd..

**Slika 8.12: Povprečna širina 100(1-α)% intervala zaupanja za presečišče**

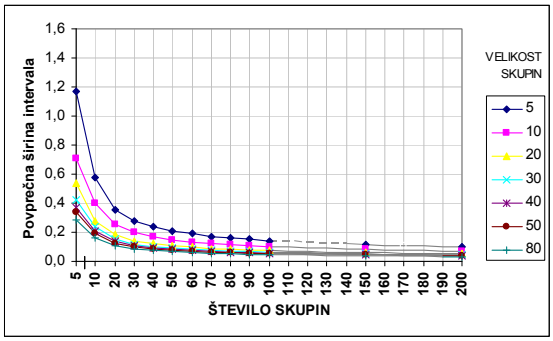



**Slika 8.13: Delež zaželene širine intervala (<0.15) glede na posamezen pogoj**

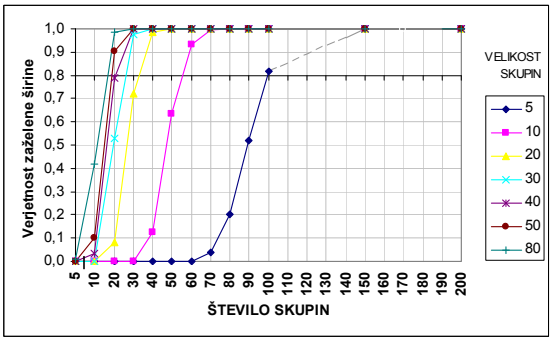


Če želimo biti v primeru nagiba $\gamma_{10} x_{ij}$ 80% prepričani, da dobimo interval zaupanja, ki ni širši od 0.15, lahko to dosežemo že z 20 skupinami z vsaj 40 enotami v skupini oz. s 30 skupinami z vsaj 30 enotami v skupini (*glej Slika 8.15*).

**Slika 8.14: Povprečna širina 100(1-α)% intervala zaupanja za nagib (X)**

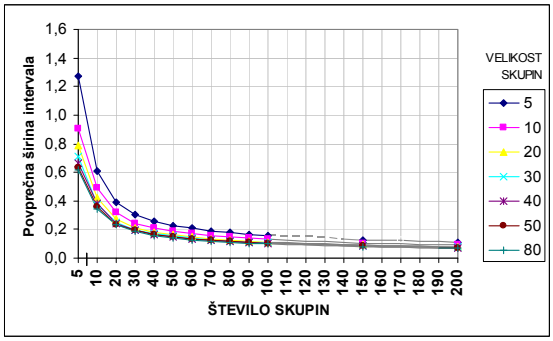



**Slika 8.15: Delež zaželene širine intervala (<0.15) glede na posamezen pogoj**

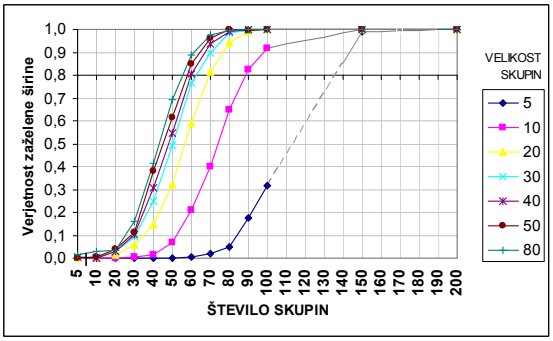



Če želimo biti v primeru $\gamma_{01} z_j$ 80% prepričani, da dobimo interval, ki ni širši od 0.15, lahko to dosežemo šele s 60 skupinami z vsaj 40 enotami v skupini oz. s 70 skupinami z vsaj 20 enotami v skupini (*glej Slika 8.17*).

**Slika 8.16: Povprečna širina 100(1-α)% intervala zaupanja za nagib (Z)**

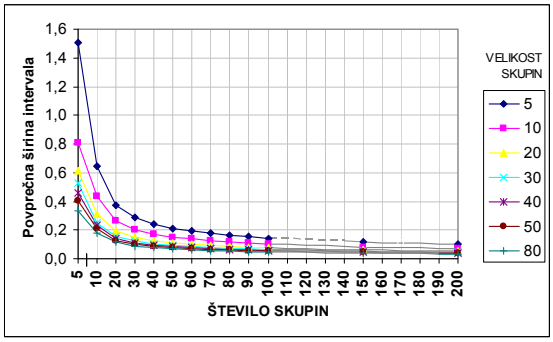



**Slika 8.17: Delež zaželene širine intervala (<0.15) glede na posamezen pogoj**

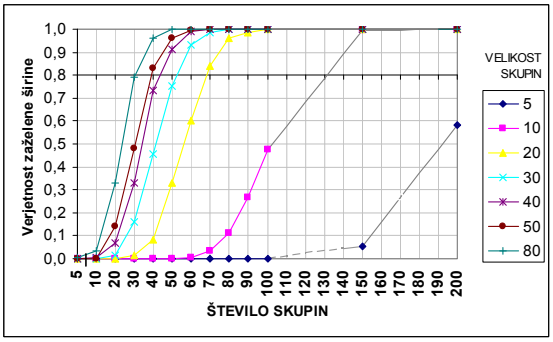



Če želimo biti v primeru $\gamma_{11} z_j x_{ij}$ 80% prepričani, da dobimo interval, ki ni širši od 0.09, lahko to dosežemo že s 40 skupinami z vsaj 50 enotami v skupini oz. s 50 skupinami z vsaj 30 enotami v skupini, itd. (*glej Slika 8.19*).

**Slika 8.18: Povprečna širina 100(1-α)% intervala zaupanja za nagib (X*Z)**



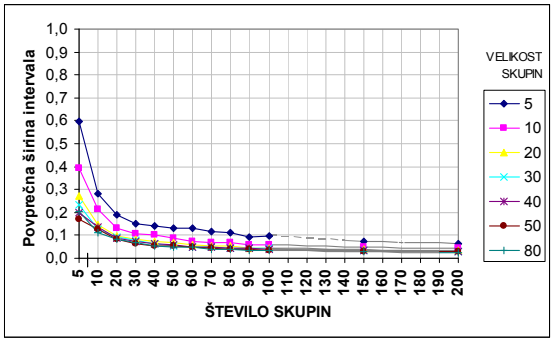

**Slika 8.19: Delež zaželene širine intervala (<0.09) glede na posamezen pogoj**



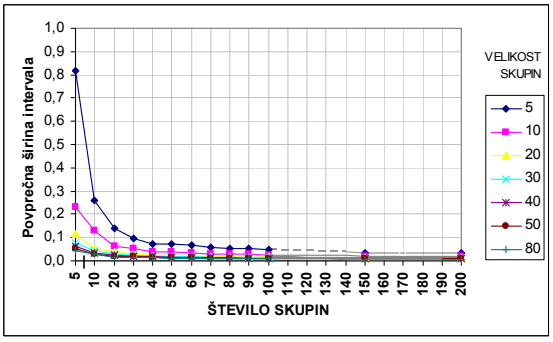
### 8.5.2 Slučajni parametri

Tudi za slučajne parametre lahko izračunamo povprečno širino intervala glede na posamezen pogoj. V primeru slučajnih parametrov so intervali zaupanja asimetrični. Lmer ne vrne standardnih napak za slučajne koeficiente, zato sem intervale zaupanja za slučajne koeficiente izračunala samo glede na posamezen pogoj (število skupin, velikost skupin), ne pa tudi za posamezno simulacijo. Intervale zaupanja za slučajne koeficiente sem tako izračunala iz posameznih empiričnih vzorčnih porazdelitev, pri čemer sem za spodnjo mejo intervala izračunala 0.025 percentil, za zgornjo mejo pa 0.975 percentil. V primeru slučajnih parametrov bom tako ostala pri pričakovani (povprečni) širini intervala. Najprej si oglejmo 95% interval zaupanja za varianco presečišča (*glej Slika 8.20*) ter za varianco nagiba (*glej Slika 8.21*).

**Slika 8.20: Povprečna širina 100(1-α)% intervala zaupanja za varianco presečišča**




**Slika 8.21: Povprečna širina 100(1-α)% intervala zaupanja za varianco nagiba**




V obeh primerih lahko opazimo močan vpliv števila skupin kot tudi velikosti skupin. V primeru variance presečišča ter variance nagiba se širina intervala opazno poveča pri številu skupin, N<30 ter pri velikostih skupin, n<20 (*glej tudi Tabela 8.7 ter Tabela 8.8*).

**Tabela 8.7: Povprečna širina 100(1-α)% intervala zaupanja za varianco presečišča**

| | VELIKOST SKUPIN | | | | | | |
|---|---|---|---|---|---|---|---|
| **ŠTEVILO SKUPIN** | **5** | **10** | **20** | **30** | **40** | **50** | **80** |
| **5** | 0,596 | 0,391 | 0,272 | 0,234 | 0,198 | 0,172 | 0,202 |
| **10** | 0,279 | 0,213 | 0,148 | 0,124 | 0,135 | 0,124 | 0,111 |
| **20** | 0,190 | 0,132 | 0,099 | 0,093 | 0,090 | 0,083 | 0,081 |
| **30** | 0,150 | 0,105 | 0,084 | 0,077 | 0,075 | 0,065 | 0,066 |
| **40** | 0,139 | 0,100 | 0,072 | 0,061 | 0,064 | 0,055 | 0,052 |
| **50** | 0,131 | 0,085 | 0,069 | 0,054 | 0,057 | 0,052 | 0,047 |
| **60** | 0,131 | 0,074 | 0,060 | 0,050 | 0,048 | 0,047 | 0,047 |
| **70** | 0,116 | 0,070 | 0,055 | 0,049 | 0,048 | 0,045 | 0,040 |
| **80** | 0,111 | 0,066 | 0,052 | 0,044 | 0,043 | 0,040 | 0,040 |
| **90** | 0,092 | 0,061 | 0,046 | 0,043 | 0,042 | 0,040 | 0,035 |
| **100** | 0,096 | 0,060 | 0,045 | 0,041 | 0,039 | 0,036 | 0,035 |
| **150** | 0,073 | 0,047 | 0,036 | 0,034 | 0,032 | 0,030 | 0,028 |
| **200** | 0,065 | 0,044 | 0,033 | 0,029 | 0,027 | 0,027 | 0,025 |

Glede na to, da gre za zelo majhne vrednosti varianc na makro nivoju, bi v primeru variance presečišča (populacijska vrednost je 0.055556) želela imeti povprečno širino

intervala največ 0.04, v primeru variance nagiba (populacijska vrednost je 0.008) pa 0.008. V tem primeru bi za testiranje variance presečišča zaželeno pričakovano širino intervala dosegla šele s 70 skupinami velikosti 80, ali 80/90 skupinami velikosti 50, 100 skupinami velikosti 40, itd. (*glej Tabela 8.7*).

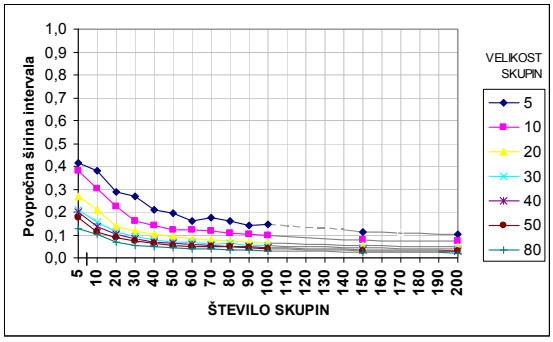
**Tabela 8.8: Povprečna širina 100(1-α)% intervala zaupanja za varianco nagiba**

| | VELIKOST SKUPIN | | | | | | |
|---|---|---|---|---|---|---|---|
| **ŠTEVILO SKUPIN** | **5** | **10** | **20** | **30** | **40** | **50** | **80** |
| **5** | 0,8182 | 0,2307 | 0,1167 | 0,0781 | 0,0618 | 0,0521 | 0,0409 |
| **10** | 0,2612 | 0,1299 | 0,0512 | 0,0427 | 0,0353 | 0,0289 | 0,0267 |
| **20** | 0,1401 | 0,0647 | 0,0382 | 0,0275 | 0,0231 | 0,0210 | 0,0168 |
| **30** | 0,0940 | 0,0518 | 0,0285 | 0,0233 | 0,0192 | 0,0174 | 0,0146 |
| **40** | 0,0723 | 0,0397 | 0,0259 | 0,0202 | 0,0172 | 0,0155 | 0,0123 |
| **50** | 0,0709 | 0,0396 | 0,0212 | 0,0183 | 0,0153 | 0,0147 | 0,0115 |
| **60** | 0,0651 | 0,0326 | 0,0203 | 0,0175 | 0,0144 | 0,0128 | 0,0103 |
| **70** | 0,0598 | 0,0305 | 0,0190 | 0,0152 | 0,0141 | 0,0119 | 0,0095 |
| **80** | 0,0534 | 0,0281 | 0,0181 | 0,0152 | 0,0127 | 0,0114 | 0,0089 |
| **90** | 0,0526 | 0,0301 | 0,0168 | 0,0144 | 0,0120 | 0,0106 | 0,0086 |
| **100** | 0,0468 | 0,0257 | 0,0177 | 0,0138 | 0,0116 | 0,0106 | 0,0078 |
| **150** | 0,0354 | 0,0207 | 0,0150 | 0,0116 | 0,0093 | 0,0082 | 0,0070 |
| **200** | 0,0331 | 0,0199 | 0,0123 | 0,0097 | 0,0083 | 0,0074 | 0,0054 |

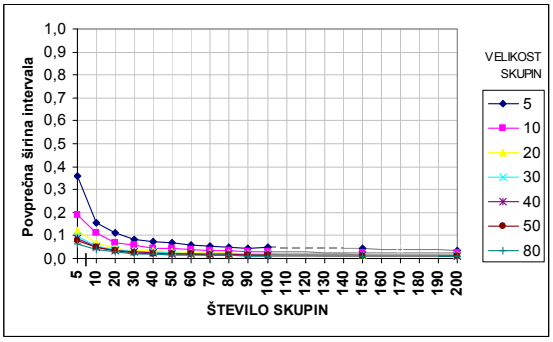
Za testiranje variance nagiba bi zaželeno pričakovano širino intervala dosegla šele s 70/80/90 skupinami velikosti 80, ali 150 skupinami velikosti 40, 200 skupinami velikosti 30 (*glej Tabela 8.8*).

Oglejmo si še 95% interval zaupanja za individualno napako (*glej Slika 8.22*) ter za kovarianco (*glej Slika 8.23*). Lahko vidimo močan vpliv števila skupin kot tudi velikosti skupin. V primeru individualne napake ima velikost skupin še močnejši vpliv kot število skupin. V primeru individualne napake velikost skupin, n<20, močno razširi interval zaupanja, glede na vse vrednosti števila skupin. V primeru kovariance večje število skupin 'popravi' širino intervala, če gre za velikost skupin 10 ali manj.

**Slika 8.22: Povprečna širina 100(1-α)% intervala zaupanja za individualno napako**




**Slika 8.23: Povprečna širina 100(1-α)% intervala zaupanja za kovarianco**



***TRETJI DEL: fokus*** Ugotovitve iz prejšnjih poglavij bom nadgradila z vključitvijo stroškovnega vidika, s pomočjo katerega bom izluščila manjše število pogojev, ki jih bom v nadaljevanju bolj podrobno preverila, prikazala tudi konkretne empirične porazdelitve, itd. in se na podlagi ugotovitev odločila za najbolj optimalno kombinacijo števila skupin ter velikosti skupin.

## 8.6 Konkretna primerjava izbranih pogojev

Izpeljava simulacij glede na vse kombinacije pogojev je zelo zamudna, zato se mora raziskovalec o kombinacijah odločiti glede na proračun, ki mu je na voljo ter okoliščine (npr. koliko je največ učencev v razredu, če želi vzorčiti učence po razredih). Denimo, da stroški za vključitev ene dodatne enote na makro nivoju v raziskavi predstavljajo vrednost, ki jo potrebujemo za opazovanje 5 enot na mikro nivoju, kar odraža razmerje v stroških (*angl. cost-ratio*) 5. S tem stroškovnim razmerjem bi za 100 skupin potrebovali enak proračun, kot bi ga potrebovali za 1000 enot v eni sami skupini, namesto 500. Za enak proračun in stroškovno razmerje bi 50 enot na makro nivoju zahtevalo vsaj 15 enot na mikro nivoju, kar pomeni 715 skupnih enot, namesto 500. Če bi vzorčili le 30 enot na makro nivoju, bi morali v povprečju vzorčiti vsaj 28 enot na mikro nivoju, kar bi pripeljalo do 840 opazovanih enot. Če bi vzorčili le 10 enot na makro nivoju, bi morali v povprečju vzorčiti 95 enot na mikro nivoju, kar bi pripeljalo do 950 opazovanih enot. Te velikosti vzorca na dveh nivojih lahko izpeljemo iz Enačbe 4.1. V Tabeli 8.9 so izračunane velikosti skupin glede na število skupin in možen proračun, ki ga ima na voljo raziskovalec.

**Tabela 8.9: Izračunane VELIKOSTI SKUPIN glede na dano število skupin in velikost proračuna**

| število skupin | PRORAČUN 1000 | | 1800 | | 2700 | |
|---|---|---|---|---|---|---|
| | velikost skupin | | | | | |
| **10** | 95 | | 175 | | 265 | |
| **20** | 45 | | 85 | | 130 | |
| **30** | 28 | 30 | 55 | 50 | 85 | 80 |
| **40** | 20 | 20 | 40 | 40 | 63 | 50 |
| **50** | 15 | 20 | 31 | 30 | 49 | 50 |
| **60** | 12 | 10 | 25 | 30 | 40 | 40 |
| **70** | 9 | 10 | 21 | 20 | 34 | 30 |
| **80** | 8 | 10 | 18 | 20 | 29 | 30 |
| **90** | 6 | 5 | 15 | 20 | 25 | 30 |
| **100** | 5 | | 13 | | 22 | |
| **110** | 4 | | 11 | | 20 | |
| **120** | 3 | | 10 | | 18 | |
| **130** | 3 | | 9 | | 16 | |
| **140** | 2 | | 8 | | 14 | |

| število skupin | PRORAČUN 1000 | 1800 | 2700 |
|---|---|---|---|
| **150** | 2 | 7 | 13 |
| **160** | 1 | 6 | 12 |
| **170** | 1 | 6 | 11 |
| **180** | 1 | 5 | 10 |
| **190** | 0 | 4 | 9 |
| **200** | 0 | 4 | 9 |

(*Obarvane vrednosti v vsakem drugem stolpcu po posameznemu proračunu predstavljajo zaokrožene vrednosti glede na moje, že obstoječe simulacije. Te vrednosti so prikazane le za število skupin, ki jih bom obravnavala v nadaljevanju.)

Denimo, da me od fiksnih parametrov zanima samo regresijski koeficient spremenljivke na makro nivoju ter regresijski koeficient za člen interakcije. Od slučajnih parametrov me zanima samo pomembnost variance slučajnega nagiba za spremenljivko na mikro nivoju. To bom ugotovila s (ponovnim) ocenjevanjem parametrov z in brez slučajnega nagiba v modelu. Izviren model je celoten model.

### 8.6.1 Primer 1: Proračun: cca.1000 enot

Denimo, da ima raziskovalec na voljo proračun za opazovanje 1000 enot. Zanima ga, kakšni bodo parametri v kombinacijah pogojev, ki jih ta omejitev prinaša. Denimo, da raziskovalec postavi omejitev, da bo raziskoval od 30 do 90 skupin, z večanjem po koraku 10. V Tabeli 8.10 so predstavljeni rezultati po posameznih pogojih. Vidimo, da bi za omenjen proračun lahko raziskovalec pričakoval precej nizko konvergenco glede na svoj model, še posebej, če namerava v posamezni skupini raziskovati 10 enot ali manj.

**Tabela 8.10: Konvergenca, natančnost ocen ter moč testa za regresijski koeficient spremenljivke na makro nivoju Z (0.10) po izbranih pogojih**

| ŠTEVILO SKUPIN | VELIKOST SKUPIN | konvergenca | povp. ocena | pristranskost | preciznost | mse | se (lmer) | sd (ESD) | moč | povp. širina 95CI | širina 95CI<0.15 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **30** | **30** | 830 | 0,09663 | -0,00337 | 0,00271 | 0,00272 | 0,05062 | 0,05205 | 0,50621 | 0,19842 | 0,09036 |
| **40** | **20** | 770 | 0,09975 | -0,00025 | 0,00200 | 0,00200 | 0,04600 | 0,0447 | 0,58477 | 0,18030 | 0,14416 |
| **50** | **20** | 818 | 0,10163 | 0,00163 | 0,00183 | 0,00183 | 0,04123 | 0,04273 | 0,67922 | 0,16162 | 0,32274 |
| **60** | **10** | 626 | 0,09748 | -0,00252 | 0,00175 | 0,00176 | 0,04319 | 0,04185 | 0,63893 | 0,16929 | 0,21086 |
| **70** | **10** | 641 | 0,10081 | 0,00081 | 0,00151 | 0,00152 | 0,03976 | 0,03892 | 0,71067 | 0,15584 | 0,40094 |
| **80** | **10** | 661 | 0,10098 | 0,00098 | 0,00146 | 0,00146 | 0,03696 | 0,03825 | 0,77201 | 0,14489 | 0,6475 |
| **90** | **5** | 508 | 0,10182 | 0,00182 | 0,00189 | 0,00189 | 0,04239 | 0,04344 | 0,65500 | 0,16618 | 0,1772 |

Poglejmo še, kaj bi iz iste tabele raziskovalec ugotovil glede natančnosti ter moči te spremenljivke. Lahko vidimo, da je pristranskost regresijskega koeficienta zelo majhna, celo zanemarljiva v vseh opazovanih pogojih. Natančnost definirana z mse je skoraj v celoti določena s preciznostjo. Standardna napaka, ki jo vrne program (*se(lmer)*) je zelo podobna standardnemu odklonu, izračunanemu iz empirične vzorčne porazdelitve (*sd(ESD)*). Največjo moč in natančnost (*širina 95CI<0.15*) ocene bi raziskovalec dobil z 80 skupinami po 10 enot, vendar ne bi dosegel želenega kriterija 0.8.

Tabela 8.11: Natančnost ocen ter moč testa za regresijski koeficient interakcijske spremenljivke po izbranih pogojih

| ŠTEVILO SKUPIN | VELIKOST SKUPIN | povp. ocena | pristranskost | preciznost | mse | se (lmer) | sd (ESD) | moč | povp. širina 95CI | širina 95CI<0.15 |
|---|---|---|---|---|---|---|---|---|---|---|
| **30** | **30** | 0,04974 | -0,00026 | 0,00103 | 0,00103 | 0,03104 | 0,03212 | 0,36341 | 0,12169 | 0,15783 |
| **40** | **20** | 0,04919 | -0,00081 | 0,00097 | 0,00097 | 0,03154 | 0,03111 | 0,35388 | 0,12365 | 0,08442 |
| **50** | **20** | 0,05065 | 0,00065 | 0,00078 | 0,00078 | 0,02742 | 0,02784 | 0,44570 | 0,10749 | 0,33007 |
| **60** | **10** | 0,04963 | -0,00037 | 0,00107 | 0,00107 | 0,03474 | 0,03264 | 0,30135 | 0,13616 | 0,00479 |
| **70** | **10** | 0,04832 | -0,00168 | 0,00097 | 0,00097 | 0,03149 | 0,03109 | 0,35486 | 0,12344 | 0,03432 |
| **80** | **10** | 0,04903 | -0,00097 | 0,00076 | 0,00076 | 0,02932 | 0,02758 | 0,39947 | 0,11494 | 0,11195 |
| **90** | **5** | 0,05298 | 0,00298 | 0,00141 | 0,00142 | 0,03868 | 0,03756 | 0,25229 | 0,15162 | 0,00000 |

Regresijski koeficient za člen interakcije (*glej Tabela 8.11*) je veliko bolj odvisen od velikosti skupin, tako da bi največjo moč ocene raziskovalec dobil s kombinacijo 50 skupin po 20 enot, vendar tudi v tem primeru ne bi dosegel 'zaželenega' kriterija 0.8. Največjo natančnost ocene bi raziskovalec prav tako dobil s kombinacijo 50 skupin po 20 enot, pri čemer tudi tu ne bi dosegel želenega kriterija 0.8.

Tabela 8.12: Natančnost ocen ter moč testa za varianco nagiba po izbranih pogojih

| ŠTEVILO SKUPIN | VELIKOST SKUPIN | povp. ocena | pristranskost | preciznost | mse | sd (ESD) |
|---|---|---|---|---|---|---|
| **30** | **30** | 0,00957 | 0,00157 | 0,00004 | 0,00004 | 0,00601 |
| **40** | **20** | 0,01104 | 0,00304 | 0,00004 | 0,00005 | 0,00669 |
| **50** | **20** | 0,00958 | 0,00158 | 0,00003 | 0,00004 | 0,00591 |
| **60** | **10** | 0,01418 | 0,00618 | 0,00008 | 0,00012 | 0,00887 |
| **70** | **10** | 0,01275 | 0,00475 | 0,00007 | 0,00010 | 0,00852 |
| **80** | **10** | 0,01222 | 0,00422 | 0,00006 | 0,00008 | 0,00781 |
| **90** | **5** | 0,01959 | 0,01159 | 0,00019 | 0,00032 | 0,01369 |

(nadaljevanje)

| ŠTEVILO SKUPIN | VELIKOST SKUPIN | povp. širina 95CI | p_Minimum | p_Median | p_Mean | p_Maximum | delež zavrnitev $H_0$ |
|---|---|---|---|---|---|---|---|
| **30** | **30** | 0,02327 | 0,00000 | 0,08385 | 0,13365 | 0,49835 | 0,39277 |
| **40** | **20** | 0,02590 | 0,00000 | 0,11082 | 0,15158 | 0,49955 | 0,34805 |
| **50** | **20** | 0,02122 | 0,00000 | 0,10054 | 0,14576 | 0,49785 | 0,34841 |
| **60** | **10** | 0,03258 | 0,00002 | 0,14970 | 0,18578 | 0,49920 | 0,22364 |
| **70** | **10** | 0,03051 | 0,00000 | 0,15794 | 0,19126 | 0,49999 | 0,22692 |
| **80** | **10** | 0,02806 | 0,00008 | 0,15576 | 0,18348 | 0,49362 | 0,24508 |
| **90** | **5** | 0,05264 | 0,00008 | 0,14370 | 0,17943 | 0,49883 | 0,25000 |

Za testiranje variance regresijskega nagiba (*glej Tabela 8.12*) bi dosegli največjo moč testa pri 30 skupinah s 30 enotami, vendar je za ta parameter moč precej nizka pri

vseh pogojih. Raziskovalec bi izgubil le malo na moči testa, če bi se tudi v tem primeru odločil za 50 skupin z 20 enotami, kjer zasledimo tudi najožji povprečni interval.

### 8.6.2 Primer 2: Proračun: cca.1800 enot

Vkolikor bi imel raziskovalec na voljo sredstva za opazovanje 1800 enot (*glej Tabela 8.13*), bi lahko pričakoval precej višjo konvergenco glede na svoj model (89%-96%), kot je bilo to v primeru proračuna za 1000 enot.

**Tabela 8.13: Konvergenca, natančnost ocen ter moč testa za regresijski koeficient spremenljivke na makro nivoju po izbranih pogojih**

| ŠTEVILO SKUPIN | VELIKOST SKUPIN | konvergenca | povp. ocena | pristranskost | preciznost | mse | se (lmer) | sd (ESD) | moč | povp. širina 95CI | širina 95CI<0.15 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **30** | **50** | 944 | 0,09969 | -0,00031 | 0,00243 | 0,00243 | 0,0491 | 0,04933 | 0,53065 | 0,19245 | 0,11441 |
| **40** | **40** | 959 | 0,10138 | 0,00138 | 0,00195 | 0,00195 | 0,04232 | 0,04419 | 0,65649 | 0,16590 | 0,30761 |
| **50** | **30** | 939 | 0,09960 | -0,00040 | 0,00156 | 0,00156 | 0,03884 | 0,03946 | 0,73072 | 0,15223 | 0,49308 |
| **60** | **30** | 943 | 0,09947 | -0,00053 | 0,00133 | 0,00133 | 0,03529 | 0,03652 | 0,80892 | 0,13832 | 0,76352 |
| **70** | **20** | 890 | 0,09881 | -0,00119 | 0,00128 | 0,00129 | 0,03455 | 0,03584 | 0,82502 | 0,13542 | 0,81573 |
| **80** | **20** | 907 | 0,10131 | 0,00131 | 0,00107 | 0,00107 | 0,03222 | 0,03270 | 0,87367 | 0,12629 | 0,94046 |
| **90** | **20** | 927 | 0,09941 | -0,00059 | 0,00088 | 0,00088 | 0,03029 | 0,02964 | 0,91016 | 0,11872 | 0,99029 |

Vidimo, da je pristranskost regresijskega koeficienta za spremenljivko na makro nivoju zanemarljiva v vseh opazovanih pogojih. Natančnost definirana z mse je skoraj v celoti določena s preciznostjo. Največjo moč in natančnost (ožji interval zaupanja) bi raziskovalec dobil z 90 skupinami po 20 enot (moč=0.91, 99% intervalov bi bilo ožjih od 0.15), vendar pa bi zadostno moč dosegel že s 60 skupinami po 30 enot, natančnost pa že s 70 skupinami po 20 enot. Optimalno moč ter natančnost bi tako dosegel s kombinacijo 70 skupin po 20 enot.

Tudi za regresijski koeficient za člen interakcije (*glej Tabela 8.14*) bi največjo moč in natančnost (ožji interval zaupanja) raziskovalec dosegel z 90 skupinami po 20 enot (moč=0.70, 98% intervalov bi bilo ožjih od 0.09), pri čemer ne bi zadostil kriteriju po želeni moči statističnega testa. Zadostno natančnost kot jo pogojuje širina intervala, bi dosegel že s kombinacijo 70 skupin po 20 enot.

Tabela 8.14: Natančnost ocen ter moč testa za regresijski koeficient interakcijske spremenljivke po izbranih pogojih

| ŠTEVILO SKUPIN | VELIKOST SKUPIN | povp. ocena | pristranskost | preciznost | mse | se (lmer) | sd (ESD) | moč | povp. širina 95CI | širina 95CI<0.15 |
|---|---|---|---|---|---|---|---|---|---|---|
| **30** | **50** | 0,05077 | 0,00077 | 0,00067 | 0,00067 | 0,02627 | 0,02594 | 0,47742 | 0,10297 | 0,48093 |
| **40** | **40** | 0,04990 | -0,00010 | 0,00063 | 0,00063 | 0,02367 | 0,02501 | 0,56061 | 0,09278 | 0,73306 |
| **50** | **30** | 0,04938 | -0,00062 | 0,00056 | 0,00056 | 0,02347 | 0,02365 | 0,56752 | 0,09202 | 0,75399 |
| **60** | **30** | 0,05143 | 0,00143 | 0,00043 | 0,00043 | 0,02119 | 0,02081 | 0,65527 | 0,08307 | 0,93425 |
| **70** | **20** | 0,05037 | 0,00037 | 0,00051 | 0,00051 | 0,02286 | 0,02251 | 0,59001 | 0,08960 | 0,84045 |
| **80** | **20** | 0,04907 | -0,00093 | 0,00044 | 0,00044 | 0,02132 | 0,02099 | 0,65006 | 0,08356 | 0,96251 |
| **90** | **20** | 0,05031 | 0,00031 | 0,00038 | 0,00038 | 0,02001 | 0,01944 | 0,70500 | 0,07844 | 0,98490 |

Za testiranje variance regresijskega nagiba (*glej Tabela 8.15*) bi dosegli največjo moč testa pri 30 skupinah s 50 enotami (0.65), najmanjšo moč pa s kombinacijo 70 skupin po 20 enot.

Tabela 8.15: Natančnost ocen ter moč testa za varianco nagiba po izbranih pogojih

| ŠTEVILO SKUPIN | VELIKOST SKUPIN | povp. ocena | pristranskost | preciznost | mse | sd (ESD) |
|---|---|---|---|---|---|---|
| **30** | **50** | 0,00856 | 0,00056 | 0,00002 | 0,00002 | 0,00455 |
| **40** | **40** | 0,00821 | 0,00021 | 0,00002 | 0,00002 | 0,00435 |
| **50** | **30** | 0,00879 | 0,00079 | 0,00002 | 0,00002 | 0,00487 |
| **60** | **30** | 0,00859 | 0,00059 | 0,00002 | 0,00002 | 0,00457 |
| **70** | **20** | 0,00864 | 0,00064 | 0,00003 | 0,00003 | 0,00500 |
| **80** | **20** | 0,00894 | 0,00094 | 0,00002 | 0,00002 | 0,00481 |
| **90** | **20** | 0,00852 | 0,00052 | 0,00002 | 0,00002 | 0,00463 |

(nadaljevanje)

| ŠTEVILO SKUPIN | VELIKOST SKUPIN | povp. širina 95CI | p_Minimum | p_Median | p_Mean | p_Maximum | delež zavrnitev $H_0$ |
|---|---|---|---|---|---|---|---|
| **30** | **50** | 0,01743 | 0,00000 | 0,01782 | 0,06946 | 0,49950 | 0,65323 |
| **40** | **40** | 0,01717 | 0,00000 | 0,02528 | 0,07506 | 0,49666 | 0,62982 |
| **50** | **30** | 0,01833 | 0,00000 | 0,03945 | 0,09659 | 0,48935 | 0,52716 |
| **60** | **30** | 0,01754 | 0,00000 | 0,02861 | 0,07569 | 0,49562 | 0,60764 |
| **70** | **20** | 0,01904 | 0,00000 | 0,07984 | 0,12712 | 0,49955 | 0,39595 |
| **80** | **20** | 0,01808 | 0,00000 | 0,05692 | 0,10732 | 0,48635 | 0,47960 |
| **90** | **20** | 0,01681 | 0,00000 | 0,05020 | 0,10269 | 0,49471 | 0,49622 |

S proračunom 1800 enot bi tako raziskovalec dosegel zadostno moč pri pogoju 90 skupin z velikostjo 20 za spremenljivko na makro nivoju, ne pa tudi za člen interakcije, kjer je moč le 0.7. Moč testa za testiranje variance nagiba je največja pri najmanjšem številu skupin (30), v kombinaciji z največjo velikostjo skupin (50). Tu je

moč 0.65. Intervali zaupanja so zadosti natančni v primeru regresijskih koeficientov, raziskovalec lahko pričakuje celo ožje intervale kot 0.15. V primeru slučajnega koeficienta je povprečni interval nekoliko širši od želenega, to je od 0.008. Če bi želel doseči nek kompromis med močjo in natančnostjo regresijskih koeficientov in varianco nagiba bi se raziskovalec lahko odločil za kombinacijo 60 skupin s 30 enotami.

### 8.6.3 Primer 3: Proračun: cca. 2700 enot

Vkolikor bi imel raziskovalec na voljo sredstva za opazovanje 2700 enot, bi lahko pričakoval zelo visoko konvergenco glede na svoj model (več kot 96%). Lahko vidimo, da je pristranskost regresijskega koeficienta za spremenljivko na makro nivoju (*glej Tabela 8.16*) zanemarljiva v vseh opazovanih pogojih. Natančnost definirana z mse je skoraj v celoti določena s preciznostjo. Največjo moč in natančnost (ožji interval zaupanja) bi raziskovalec dobil z 90 skupinami po 30 enot (moč=0.94, 99.7% intervalov bi bilo ožjih od 0.15), vendar pa bi zadostno moč in natančnost dosegel že s 60 skupinami po 40 enot.

**Tabela 8.16: Konvergenca, natančnost ocen ter moč testa za regresijski koeficient spremenljivke na makro nivoju po izbranih pogojih**

| ŠTEVILO SKUPIN | VELIKOST SKUPIN | konvergenca | povp. ocena | pristranskost | preciznost | mse | se (lmer) | sd (ESD) | moč | povp. širina 95CI | širina 95CI<0.15 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **30** | **80** | 988 | 0,10190 | 0,00190 | 0,00211 | 0,00212 | 0,04700 | 0,04597 | 0,56663 | 0,18423 | 0,16194 |
| **40** | **50** | 969 | 0,10033 | 0,00033 | 0,00162 | 0,00162 | 0,04104 | 0,04023 | 0,68331 | 0,16086 | 0,37874 |
| **50** | **50** | 983 | 0,09856 | -0,00144 | 0,00135 | 0,00135 | 0,03707 | 0,03677 | 0,76953 | 0,14533 | 0,61343 |
| **60** | **40** | 989 | 0,10101 | 0,00101 | 0,00119 | 0,00119 | 0,03438 | 0,0345 | 0,82851 | 0,13479 | 0,80586 |
| **70** | **30** | 964 | 0,09935 | -0,00065 | 0,00108 | 0,00108 | 0,03266 | 0,03291 | 0,86480 | 0,12801 | 0,89834 |
| **80** | **30** | 980 | 0,10160 | 0,00160 | 0,00092 | 0,00093 | 0,03044 | 0,03038 | 0,90743 | 0,11933 | 0,98367 |
| **90** | **30** | 986 | 0,10014 | 0,00014 | 0,00082 | 0,00082 | 0,02875 | 0,02856 | 0,93559 | 0,11268 | 0,99696 |

Tudi za regresijski koeficient za člen interakcije (*glej Tabela 8.17*) bi največjo (in zadostno) moč in natančnost (ožji interval zaupanja) raziskovalec dosegel z 90 skupinami po 30 enot (moč=83.7, 100% intervalov bi bilo ožjih od 0.09). Zadostno natančnost kot jo pogojuje širina intervala, bi dosegel že s kombinacijo 40 skupin po 50 enot (0.83) (oz. če nismo preveč strogi že s 30 skupinami po 80 enot (0.79)).

Tabela 8.17: Natančnost ocen ter moč testa za regresijski koeficient interakcijske spremenljivke po izbranih pogojih

| ŠTEVILO SKUPIN | VELIKOST SKUPIN | povp. ocena | pristranskost | preciznost | mse | se (lmer) | sd (ESD) | moč | povp. širina 95CI | širina 95CI<0.15 |
|---|---|---|---|---|---|---|---|---|---|---|
| **30** | **80** | 0,05142 | 0,00142 | 0,00048 | 0,00048 | 0,02253 | 0,02188 | 0,60233 | 0,08831 | 0,79251 |
| **40** | **50** | 0,04950 | -0,00050 | 0,00048 | 0,00048 | 0,02214 | 0,02186 | 0,61727 | 0,08679 | 0,82972 |
| **50** | **50** | 0,05009 | 0,00009 | 0,00041 | 0,00041 | 0,01971 | 0,02036 | 0,71793 | 0,07727 | 0,96338 |
| **60** | **40** | 0,05025 | 0,00025 | 0,00037 | 0,00037 | 0,01914 | 0,01926 | 0,74285 | 0,07503 | 0,98989 |
| **70** | **30** | 0,04937 | -0,00063 | 0,00037 | 0,00037 | 0,01935 | 0,01929 | 0,73351 | 0,07587 | 0,98651 |
| **80** | **30** | 0,05043 | 0,00043 | 0,00035 | 0,00035 | 0,01817 | 0,01864 | 0,78561 | 0,07124 | 0,99796 |
| **90** | **30** | 0,04990 | -0,00010 | 0,00029 | 0,00029 | 0,01700 | 0,01702 | 0,83684 | 0,06663 | 1,00000 |

Za testiranje variance regresijskega nagiba (*glej Tabela 8.18*) bi dosegli največjo moč testa pri 30 skupinah z 80 enotami (0.86), zadostno moč pa bi dosegli tudi s 50 skupinami po 50 enot, tej vrednosti bi se približalo tudi 60 skupin s 40 enotami (78%). Najmanjšo standardno napako zabeležimo pri pogoju 60 skupin s 40 enotami ter pri pogoju 90 skupin s po 30 enotami.

Tabela 8.18: Natančnost ocen ter moč testa za varianco nagiba po izbranih pogojih

| ŠTEVILO SKUPIN | VELIKOST SKUPIN | povp. ocena | pristranskost | preciznost | mse | sd (ESD) |
|---|---|---|---|---|---|---|
| **30** | **80** | 0,00804 | 0,00004 | 0,00001 | 0,00001 | 0,00380 |
| **40** | **50** | 0,00819 | 0,00019 | 0,00002 | 0,00002 | 0,00397 |
| **50** | **50** | 0,00827 | 0,00027 | 0,00001 | 0,00001 | 0,00385 |
| **60** | **40** | 0,00809 | 0,00009 | 0,00001 | 0,00001 | 0,00369 |
| **70** | **30** | 0,00806 | 0,00006 | 0,00002 | 0,00002 | 0,00399 |
| **80** | **30** | 0,00819 | 0,00019 | 0,00002 | 0,00002 | 0,00398 |
| **90** | **30** | 0,00798 | -0,00002 | 0,00001 | 0,00001 | 0,00373 |

(nadaljevanje)

| ŠTEVILO SKUPIN | VELIKOST SKUPIN | povp. širina 95CI | p_Minimum | p_Median | p_Mean | p_Maximum | delež zavrnitev $H_0$ |
|---|---|---|---|---|---|---|---|
| **30** | **80** | 0,01457 | 0 | 0,00098 | 0,02552 | 0,48874 | 0,8583 |
| **40** | **50** | 0,01547 | 0 | 0,00821 | 0,04473 | 0,47611 | 0,76319 |
| **50** | **50** | 0,01473 | 0 | 0,00217 | 0,03332 | 0,49445 | 0,81837 |
| **60** | **40** | 0,01442 | 0 | 0,00665 | 0,04227 | 0,46807 | 0,77856 |
| **70** | **30** | 0,01517 | 0 | 0,02065 | 0,0693 | 0,49906 | 0,65145 |
| **80** | **30** | 0,01517 | 0 | 0,01234 | 0,05986 | 0,49548 | 0,70714 |
| **90** | **30** | 0,01436 | 0 | 0,0083 | 0,05155 | 0,49395 | 0,74568 |

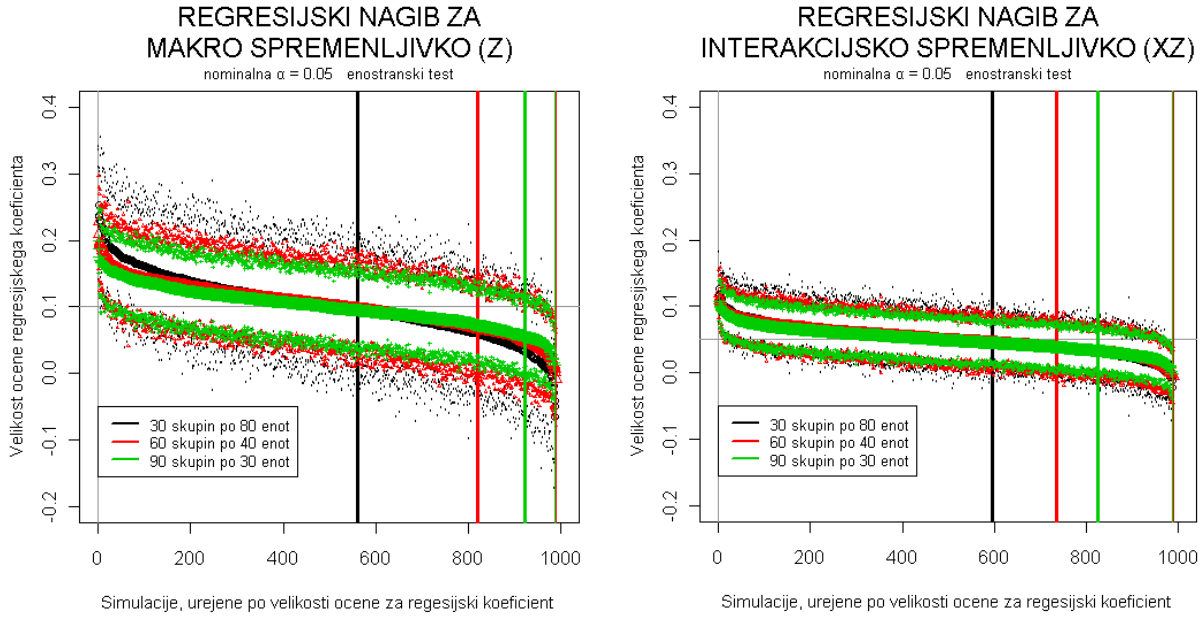
Če bi želel doseči nek kompromis med močjo in natančnostjo regresijskih koeficientov in varianco nagiba bi se raziskovalec odločil za kombinacijo 90 skupin s

30 enotami, kjer bi pri vseh preučevanih parametih dosegel zadostno moč ter natančnost. Denimo, da bi raziskovalec želel podrobneje preveriti obnašanje preučevanih parametrov glede na 3 pogoje, ki so se izkazali za najbolj optimalne (30/80,60/40,90/30) glede na tri možne proračune. V ta namen si bomo v naslednjem poglavju pogledali še celotne empirične porazdelitve teh parametrov.

## 8.7 Empirične vzorčne porazdelitve po izbranih pogojih

V tem delu bom poleg opisnih statistik, ki smo jih obravnavali do sedaj, lahko predstavila tudi urejeno množico ocen in njihovih standardnih napak za vsakega od omenjenih pogojev. Na ta način lahko primerjamo statistike tudi grafično in v obzir namesto zgolj opisnih statistik vzamemo celotno porazdelitev (empirično vzorčno porazdelitev). Za teste največjega verjetja lahko za vsak pogoj v graf izrišemo tudi p-vrednosti. Na Sliki 8.24 sta za izbrana regresijska koeficienta prikazana grafa, ki prikazujeta urejeno množico veljavnih ocen in njihove intervale zaupanja za naslednje pogoje (90 skupin po 30 enot, 60 skupin po 40 enot ter 30 skupin po 80 enot). Preciznost ocene je prikazana s spodnjo in zgornjo mejo intervalov zaupanja, ki jih prikazujejo točke. Lahko vidimo, da je pri pogoju s samo 30 skupinami preciznost ocen v splošnem manjša (intervali zaupanja so večji) in da veliko bolj variira po posameznih vzorcih zaradi omejenega števila enot, iz katerih je izračunana varianca. Nasprotno pa je preciznost največja v primeru, ko imamo največ skupin (90 skupin po 30 enot). Preciznost je večja v primeru interakcijske spremenljivke.

**Slika 8.24: Urejena množica veljavnih ocen in njihovih intervalov zaupanja, s spodnjo in zgornjo mejo, za spremenljivko na makro nivoju (b=0.1) (levo) ter za interakcijsko spremenljivko (b=0.05) (desno) za tri izbrane pogoje (različne barve). Dodatno so vključene vertikalne črte, ki ponazarjajo število zavrnitev $H_0$ (interval ne vključuje ničle: debele črte) ter število veljavnih oz. konvergiranih ocen (tanke črte).**




Preciznost ocenjevanja pa lahko nadalje aproksimiramo z vertikalno razširjenostjo ocen (ki jo predstavlja polna krivulja), ki kaže na to, da je pogoj s 30 skupinami manj

precizen. To lahko vidimo tudi iz debelejših vertikalnih črt, ki kažejo na stopnjo zavrnitev ničelne hipoteze. Tanke črte kažejo na število veljavnih ocen in jih lahko primerjamo s 1000 simulacijami, ki smo jih želeli. Lahko vidimo, da imamo zaradi nekonvergence zelo majhno izgubo v vseh treh primerih.

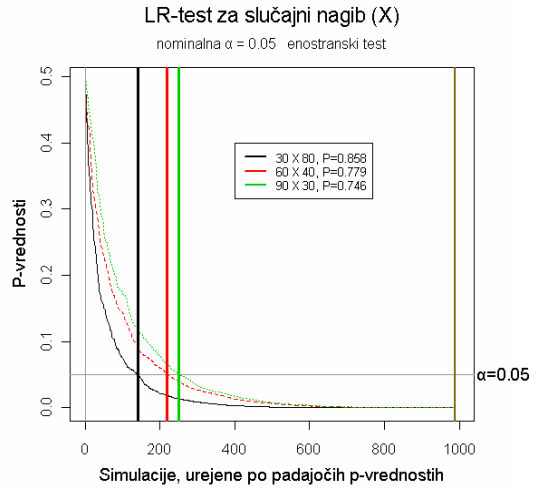
Za testiranje varianc na višjih nivojih postane pristranskost pogosto resen problem, če gre za manjše število enot, zato so v teh primerih vzorčne porazdelitve samo zelo na grobo aproksimirane z normalno porazdelitvijo. Zaradi tega je priporočljivo namesto Waldovega testa izpeljati test največjega verjetja (LR test). Še več, poleg samih ocen v nadaljevanju prikazujemo še rezultate LR testa ter njihove p-vrednosti. Če izberemo stopnjo značilnosti α=0.05, potem Tabela 8.19 kaže število zavrnitev ničelne hipoteze in še nekatere statistike za porazdelitev p-vrednosti, ki izhajajo iz LR testa z mešano hi-kvadrat porazdelitvijo.

**Tabela 8.19: Povzetek rezultatov LR testa za varianco slučajnega nagiba (u1) za tri pogoje (30/60/90 skupin) vključujoč število zavrnitev ničelne hipoteze, populacijsko vrednost variance slučajnega nagiba in nekatere osnovne statistike**

| N | n | delež zavrnitev $H_0$ | velikost vpliva | Minimum | Mediana | Povprečje | Maksimum |
|---|---|---|---|---|---|---|---|
| 30 | 80 | 0,85830 | 0,008 | 1,38778E-14 | 0,00098 | 0,02552 | 0,48874 |
| 60 | 40 | 0,77856 | 0,008 | 4,09199E-10 | 0,00665 | 0,04227 | 0,46807 |
| 90 | 30 | 0,74568 | 0,008 | 3,96848E-11 | 0,00830 | 0,05155 | 0,49395 |

Rezultate lahko prikažemo tudi v sliki (*glej Slika 8.25*), pri čemer ponovno izberemo nominalno stopnjo značilnosti, α=0.05. Pri določenih pogojih ničelno hipotezo, da ni variance nagiba spremenljivke na mikro nivoju, zavrnemo v 75%-86% ponovitvah, odvisno od pogoja.

**Slika 8.25: Urejena množica veljavnih ocen p-vrednosti za testiranje slučajnega nagiba (X) za tri pogoje: vzorčenih 30, 60 ali 90 skupin. Nominalna alfa (horizontalna linija) seka število zavrnitev (vertikalna linija).**




Vidimo, da največjo moč za testiranje variance nagiba dosežemo, ko imamo vzorčenih najmanj skupin, a največ enot na mikro nivoju (*glej Slika 8.25*). Naj na tem mestu ponovim, da prikazane p-vrednosti izhajajo iz enako uteženih hi-kvadrat porazdelitev z 0 in 2 prostostnima stopnjama.

***ČETRTI DEL: analiza občutljivosti*** Denimo, da se raziskovalec odloči za optimalen vzorec 60 skupin s 40 enotami, kljub temu, da je pri tem pogoju moč za regresijski koeficient za spremenljivko na makro nivoju 'le' 0.74. Vseeno pa bi raziskovalec želel preveriti še občutljivost parametrov v njegovem modelu s t.i. analizo občutljivosti (*angl. sensitivity analysis*). Konkretneje ga zanima, do kakšnih sprememb bo prišlo v primeru neuravnoteženih podatkov (predvideva npr. 20% upad respondentov, manjkajoče vrednosti,…), oz. če bi bili fiksni oz. slučajni parametri večji oz. manjši od predvidene velikosti vpliva.

## 8.8 Občutljivost na neuravnoteženost podatkov

Kompleksnost večnivojskih modelov dodatno povečajo neuravnoteženi dizajni, ki predstavljajo dodaten razlog, da se izvedejo analize natančnosti ter moči. Raven do katere je uravnoteženost izkrivljena lahko variira.[25] V pričujoči nalogi bom primerjala uravnotežene podatkovne strukture s podatkovnimi strukturami z različnimi velikostmi skupin. Ugotavljala bom vpliv dveh vrst neuravnoteženega dizajna: (i) neuravnoteženost v smislu velikosti odklona od povprečnega števila skupin ter (ii) neuravnoteženost v smislu povečevanja števila majhnih skupin.

### 8.8.1 Varianta A: Neuravnoteženost v smislu odklona od povprečnega števila skupin

Denimo, da raziskovalca zanima kako bi na rezultate vplivalo, če bi bile velikosti skupin neuravnotežene na naslednje načine. Velikosti skupine se od povprečne velikosti (40) razlikujejo s standardnimi odkloni 5, 10 ter 20. V Tabeli 8.20 so prikazane velikosti skupin glede na posamezne standardne odklone.

**Tabela 8.20: Prikaz dizajnov glede na tri nivoje neuravnoteženosti**

| Standardni odklon | Velikosti skupin | Skupno število enot |
|---|---|---|
| **SD=5** | 36 41 48 34 40 41 44 39 50 39 42 45 38 35 49 28 44 40 45 42 50 34 48 50 40 28 42 37 44 41 44 42 45 39 36 37 31 35 37 39 38 30 36 50 43 50 38 40 39 34 36 50 37 46 35 30 38 45 46 48 | 2428 |
| **SD=10** | 22 60 33 42 45 32 20 35 41 31 31 43 39 44 39 31 53 48 51 26 50 23 35 26 18 58 33 37 36 44 56 57 28 26 25 27 60 40 32 34 51 43 37 33 31 60 49 60 36 36 30 37 45 54 46 45 52 51 41 32 | 2380 |
| **SD=20** | 65 43 74 31 19 51 27 53 6 5 54 47 57 2 64 64 61 56 82 11 28 48 24 42 55 27 53 51 24 20 60 37 54 23 66 13 55 49 45 53 48 27 35 47 14 22 82 2 15 60 62 57 41 46 22 27 35 21 56 8 | 2426 |

[25] Uravnoteženost tako lahko npr. pokvarimo tako, da predvidevamo, da bo 50 od 100 skupin vsebovalo 50 enot, druga polovica pa 150. Večjo izkrivljenost lahko predpostavljamo, če imamo npr. 10 enot v polovici skupin in 190 v drugi polovici. Drugo; uravnoteženost je lahko pokvarjena, če se samo relativno majhen delež skupin značilno razlikuje od velikosti ostalih skupin. Primer je podatkovna struktura z 90 skupinami z 110 enotami in 10 skupin z 10 enotami, kar nam da povprečno velikost skupine 100.

***Spremenljivka na makro nivoju*** Iz Tabele 8.21 lahko vidimo, da natančnost koeficienta v smislu mse pada s povečanjem neuravnoteženosti velikosti skupin, vendar pa so te razlike minimalne. Neuravnoteženost nima znatnega vpliva na (ne)konvergenco.

**Tabela 8.21: Konvergenca in regresijski koeficient za spremenljivko na makro nivoju (Z)**

| št. skupin=60, velikost skupin | iter | povprečje | pristranskost | preciznost | mse | se (lmer) | sd (ESD) | moč | povp. širina 95CI | širina 95CI<0.15 |
|---|---|---|---|---|---|---|---|---|---|---|
| **40** | 989 | 0,10101 | 0,00101 | 0,00119 | 0,00119 | 0,03438 | 0,03450 | 0,82851 | 0,13479 | 0,80586 |
| **40±5** | 989 | 0,09968 | -0,00032 | 0,00120 | 0,00120 | 0,03417 | 0,03466 | 0,83305 | 0,13396 | 0,83721 |
| **40±10** | 991 | 0,10034 | 0,00034 | 0,00121 | 0,00121 | 0,03455 | 0,03475 | 0,82505 | 0,13542 | 0,79213 |
| **40±20** | 992 | 0,10068 | 0,00068 | 0,00130 | 0,00130 | 0,03557 | 0,03600 | 0,80269 | 0,13944 | 0,73891 |

S povečevanjem neuravnoteženosti skupin povečujemo standardno napako. Povečanje standardne napake zaradi neuravnoteženosti vpliva na moč ter širino intervala. Moč testa se pri največji neuravnoteženosti zmanjša iz 0.83 na 0.80, verjetnost, da bo interval manjši od zaželene širine (0.09), pa iz 0.81 na 0.74. Neuravnoteženost v našem primeru tako v največji meri vpliva na verjetnost zaželene širine intervala, ki pade nekoliko pod zaželeno mejo, na 0.74.

***Člen interakcije*** Iz Tabele 8.22 vidimo, da natančnost koeficienta v smislu mse pada s povečanjem neuravnoteženosti velikosti skupin, vendar pa so te razlike minimalne in se kažejo šele na 5 decimalnem mestu. S povečanjem neuravnoteženosti skupin povečujemo tudi standardno napako. Razlike se kažejo šele na četrtem decimalnem mestu.

**Tabela 8.22: Regresijski koeficient za člen interakcije (XZ)**

| št. skupin=60, velikost skupin | povprečje | pristranskost | preciznost | mse | se (lmer) | sd (ESD) | moč | povp. širina 95CI | širina 95CI<0.15 |
|---|---|---|---|---|---|---|---|---|---|
| **40** | 0,05025 | 0,00025 | 0,00037 | 0,00037 | 0,01914 | 0,01926 | 0,74285 | 0,07503 | 0,98989 |
| **40±5** | 0,04962 | -0,00038 | 0,00038 | 0,00038 | 0,01907 | 0,01940 | 0,74604 | 0,07475 | 0,92800 |
| **40±10** | 0,05035 | 0,00035 | 0,00039 | 0,00039 | 0,01935 | 0,01966 | 0,73374 | 0,07585 | 0,90700 |
| **40±20** | 0,04952 | -0,00048 | 0,00039 | 0,00039 | 0,01973 | 0,01984 | 0,71703 | 0,07735 | 0,88200 |

Povečanje standardne napake zaradi neuravnoteženosti vpliva na moč ter širino intervala. Moč testa se pri največji neuravnoteženosti zmanjša iz 0.74 na 0.72, verjetnost, da bo interval manjši od zaželene širine (0.09), pa iz 0.99 na 0.88. Posledica neuravnoteženosti se torej v največji meri kaže na verjetnosti želene širine intervala, vendar le-ta ne pade pod želeno mejo, 0.8.

***Varianca nagiba*** Iz Tabele 8.23 lahko vidimo, da natančnost koeficienta v smislu mse pada s povečanjem neuravnoteženosti velikosti skupin, vendar pa so te razlike minimalne in se kažejo šele na 5 decimalnem mestu (še to samo v primeru največje neuravnoteženosti). S povečanjem neuravnoteženosti skupin povečujemo tudi standardno napako, še posebej, če gre za večjo neuravnoteženost (SD±20).

**Tabela 8.23: Varianca nagiba (u0)**

| št. skupin=60, velikost skupin→ | povprečje | pristranskost | preciznost | mse | SE: sd (ESD) | povp. širina 95CI |
|---|---|---|---|---|---|---|
| **40** | 0,00809 | 0,00009 | 0,00001 | 0,00001 | 0,00369 | 0,01442 |
| **40±5** | 0,00814 | 0,00014 | 0,00001 | 0,00001 | 0,00381 | 0,01488 |
| **40±10** | 0,00813 | 0,00013 | 0,00001 | 0,00001 | 0,00380 | 0,01475 |
| **40±20** | 0,00796 | -0,00004 | 0,00002 | 0,00002 | 0,00395 | 0,01581 |

**(nadaljevanje)**

| št. skupin=60, velikost skupin→ | p_Minimum | p_Median | p_Mean | p_Maximum | moč (delež zavrnitev H0) |
|---|---|---|---|---|---|
| **40** | 0 | 0,00665 | 0,04227 | 0,46807 | 0,77856 |
| **40±5** | 0 | 0,00556 | 0,04089 | 0,47934 | 0,79070 |
| **40±10** | 0 | 0,00510 | 0,04348 | 0,49826 | 0,77800 |
| **40±20** | 0 | 0,00417 | 0,04142 | 0,47234 | 0,79032 |

Neuravnoteženost nima jasnega vpliva na moč testa za varianco nagiba. Moč testa variira od 0.78 do 0.79.

### 8.8.2 Varianta B: Neuravnoteženost v smislu deleža majhnih skupin

Denimo, da raziskovalec predvidi, da bo v določenem deležu skupin mogoče anketirati le 5 enot. Odstotki skupin s 5 enotami bodo variirali na naslednje načine (*glej Tabela 8.24*):

**Tabela 8.24: Prikaz dizajnov glede na tri nivoje neuravnoteženosti**

| Procent majhnih skupin | Načrt skupin | | Skupno število enot |
|---|---|---|---|
| | Število majhnih skupin z velikostjo n=5 | Število skupin z velikostjo n=40 | |
| 5% | 3 | 57 | 2295 |
| 10% | 6 | 54 | 2190 |
| 15% | 9 | 51 | 2085 |
| 20% | 12 | 48 | 1980 |

***Spremenljivka na makro nivoju*** Iz Tabele 8.25 lahko vidimo, da natančnost koeficienta v smislu mse pada s povečanjem deleža majhnih skupin, vendar pa so te razlike minimalne.

**Tabela 8.25: Konvergenca in regresijski koeficient za spremenljivko na makro nivoju (Z)**

| N=60, % majhnih skupin (n=5) | iter | povprečje | pristranskost | preciznost | mse | se (lmer) | sd (ESD) | moč | širina povp. 95CI | širina 95CI<0.15 |
|---|---|---|---|---|---|---|---|---|---|---|
| **0%** | 989 | 0,10101 | 0,00101 | 0,00119 | 0,00119 | 0,03438 | 0,03450 | 0,82851 | 0,13479 | 0,80586 |
| **5%** | 996 | 0,09979 | -0,00021 | 0,00124 | 0,00124 | 0,03477 | 0,03523 | 0,82025 | 0,13628 | 0,77510 |
| **10%** | 993 | 0,09879 | -0,00121 | 0,00126 | 0,00126 | 0,03547 | 0,03548 | 0,80494 | 0,13904 | 0,72105 |
| **15%** | 991 | 0,10011 | 0,00011 | 0,00132 | 0,00132 | 0,03586 | 0,03634 | 0,79628 | 0,14058 | 0,71645 |
| **20%** | 985 | 0,09885 | -0,00115 | 0,00138 | 0,00139 | 0,03629 | 0,03721 | 0,78690 | 0,14225 | 0,67513 |

S povečanjem deleža majhnih skupin povečujemo tudi standardno napako. Standardna napaka se poveča iz 0.03450 na 0.03721, kar vodi do zmanjšanja moči iz 0.83 na 0.79. Verjetnost zaželene širine intervala se zmanjša iz 0.81 na 0.67. V tem smislu moč testa ne predstavlja problema, ga pa lahko predstavlja verjetnost zaželene širine intervala, ki pade kar precej pod optimalno mejo, 0.8.

***Člen interakcije*** Iz Tabele 8.26 lahko vidimo, da natančnost koeficienta v smislu mse pada s povečanjem neuravnoteženosti velikosti skupin, vendar pa so te razlike minimalne in se kažejo šele na 4 decimalnem mestu. S povečanjem neuravnoteženosti skupin povečujemo tudi standardno napako.

**Tabela 8.26: Regresijski koeficient za člen interakcije**

| N=60, % majhnih skupin (n=5) | povprečje | pristranskost | preciznost | mse | se (lmer) | sd (ESD) | moč | povp. širina 95CI | širina 95CI<0.15 |
|---|---|---|---|---|---|---|---|---|---|
| **0%** | 0,05025 | 0,00025 | 0,00037 | 0,00037 | 0,01914 | 0,01926 | 0,74285 | 0,07503 | 0,98989 |
| **5%** | 0,05085 | 0,00085 | 0,00038 | 0,00039 | 0,01945 | 0,01962 | 0,72912 | 0,07626 | 0,88755 |
| **10%** | 0,05002 | 0,00002 | 0,00041 | 0,00041 | 0,01988 | 0,02021 | 0,71059 | 0,07793 | 0,86405 |
| **15%** | 0,04916 | -0,00084 | 0,00043 | 0,00043 | 0,02049 | 0,02066 | 0,68438 | 0,08033 | 0,80727 |
| **20%** | 0,05005 | 0,00005 | 0,00044 | 0,00044 | 0,02088 | 0,02094 | 0,66829 | 0,08183 | 0,77970 |

Moč testa se pri največji neuravnoteženosti zmanjša iz 0.74 na 0.67. Verjetnost zaželene širine intervala se zmanjša iz 0.99 na 0.78. Kljub neuravnoteženosti torej lahko pričakujemo dovolj natančno oceno iz vidika širine intervala, medtem ko je moč testa nekoliko bolj problematična.

***Varianca nagiba*** Iz Tabele 8.27 lahko vidimo, da natančnost koeficienta v smislu mse pada s povečanjem neuravnoteženosti velikosti skupin, vendar pa so te razlike minimalne in se kažejo šele na 5 decimalnem mestu. S povečanjem neuravnoteženosti skupin povečujemo tudi standardno napako, še posebej, če gre za večjo neuravnoteženost (20% skupin ima le 5 enot).

**Tabela 8.27: Regresijski koeficient za varianco nagiba**

| N=60, % majhnih skupin (n=5) | povprečje | pristranskost | preciznost | mse | sd (ESD) | povp. širina 95CI |
|---|---|---|---|---|---|---|
| **0%** | 0,00809 | 0,00009 | 0,00001 | 0,00001 | 0,00369 | 0,01442 |
| **5%** | 0,00793 | -0,00007 | 0,00002 | 0,00002 | 0,00402 | 0,01554 |
| **10%** | 0,00803 | 0,00003 | 0,00002 | 0,00002 | 0,00402 | 0,01541 |
| **15%** | 0,00825 | 0,00025 | 0,00002 | 0,00002 | 0,00395 | 0,01505 |
| **20%** | 0,00813 | 0,00013 | 0,00002 | 0,00002 | 0,00418 | 0,01662 |

Tabela 8.28: Regresijski koeficient za varianco nagiba (nadaljevanje tabele)

| N=60, % majhnih skupin (n=5) | p_Minimum | p_Median | p_Mean | p_Maximum | moč (delež zavrnitev H0) |
|---|---|---|---|---|---|
| **0%** | 0 | 0,00665 | 0,04227 | 0,46807 | 0,77856 |
| **5%** | 0 | 0,05334 | 0,00958 | 0,49386 | 0,72289 |
| **10%** | 0 | 0,05337 | 0,00954 | 0,49715 | 0,73414 |
| **15%** | 0 | 0,05282 | 0,00955 | 0,49465 | 0,73360 |
| **20%** | 0 | 0,06168 | 0,01584 | 0,49735 | 0,68629 |

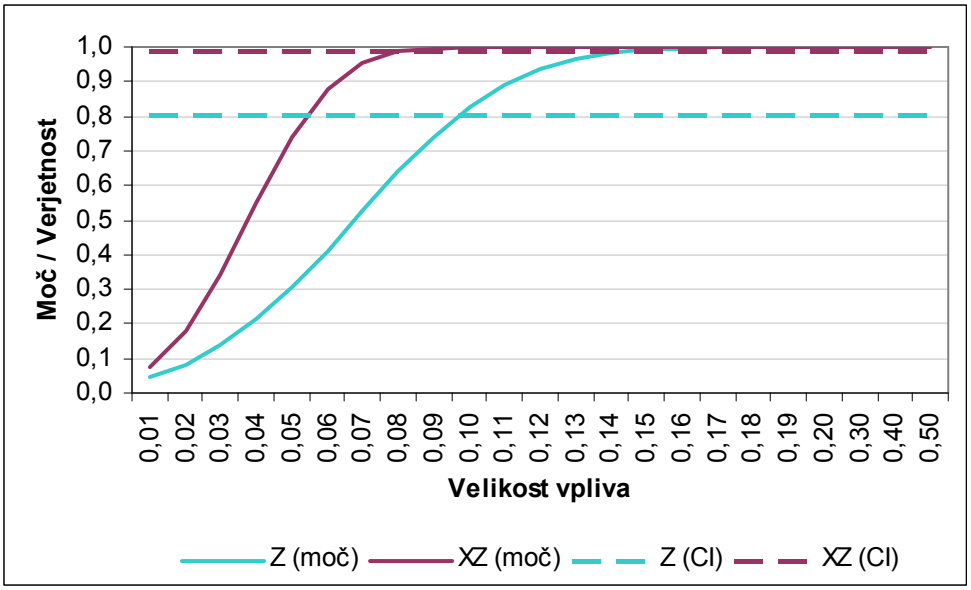
Tudi v tem primeru neuravnoteženost nima jasnega vpliva na moč testa za varianco nagiba, dokler skupin z velikostjo 5 v raziskovalnem načrtu ni več kot 20%.

## 8.9 Občutljivost na spreminjanje velikosti vpliva

Ko načrtujemo primerno velikost vzorca, ne glede na to, ali gre za perspektivo PA ali AIPE, je ponavadi nerealistično pričakovati, da so vrednosti potrebnih populacijskih parametrov natančno znane. Zato je zaželeno, da raziskovalec, ki uporablja metode načrtovanja velikosti vzorca, izvede t.i analizo občutljivosti (*angl. sensitivity analysis*) tudi glede na različne vrednosti parametrov. Le-ta vključuje izračunavanje potrebne velikosti vzorca z uporabo serije realističnih vrednosti potrebnih populacijskih parametrov. Denimo, da raziskovalca zanima, kakšen je vpliv spreminjanja velikosti regresijskih koeficientov spremenljivke na makro nivoju ter vpliva interakcije na moč. Naj ponovno spomnim bralca, da spreminjanje velikosti regresijskih koeficientov nima vpliva na preciznost, to je širino intervala (*glej konstantne vodoravne črte v Sliki 8.26*).

V Sliki 8.26 so prikazane krivulje za moč ter verjetnost, da naš interval zaupanja ne bo širši od želene širine. Lahko vidimo kako se moč spreminja glede na zmanjšanje velikosti vpliva po koraku 0.01 do 0.5. Vidimo, da je moč testa močno odvisna od velikosti vpliva in da bi, v primeru, da je raziskovalec podcenil dejansko velikost vpliva, to pripeljalo do znatno nižje moči testa kot je to potrebno.

**Slika 8.26: Prikaz občutljivosti moči statističnega testa (moč) in zaželene verjetnosti za širino intervala (CI) za regresijski nagib spremenljivke na makro nivoju (Z) ter za regresijski nagib interakcijske spremenljivke (XZ) glede na različne vrednosti velikosti vpliva**



# 9 Zaključek

V pričujoči magistrski nalogi sem raziskovala kako je učinkovitost načrta večnivojske raziskave odvisna od vzorčenja na prvem in drugem nivoju, pri čemer sem preverjala tudi vpliv različnih dejavnikov, kot so različne porazdelitve varianc med nivoji (različni znotrajrazredni koeficienti), neuravnoteženost vzorcev, občutljivost natančnosti ter moči glede na velikost vpliva. Poleg vsega naštetega sem v analizo vključila tudi stroškovni vidik.

Majhni vzorci so se izkazali za problematične že na samem začetku analize, saj algoritem ni vedno skonvergiral. Za opazno izboljšanje konvergence potrebujemo vsaj 20-30 skupin, v primeru manjših velikosti skupin in/ali znotrajrazrednih koeficientov pa še več. Konvergenca je izraziteje slabša, ko skupine vsebujejo le 5 ali 10 enot. V primeru ***regresijskih koeficientov*** z večjo težo (vsaj srednja velikost vpliva, B≥0.3) je pristranskost zanemarljiva tudi v primeru manjših vzorcev. Ko gre za manjše velikosti vplivov, je priporočljivo imeti v vzorcu vsaj 20 skupin, z vsaj 10 enotami v skupini. Znotrajrazredni koeficient na pristranskost regresijskih koeficientov nima vpliva. Standardna napaka regresijskih koeficientov se povečuje z večanjem znotrajrazrednega korelacijskega koeficienta ter z zmanjševanjem števila skupin (N<20) ter velikosti skupin (n<10). Standardne napake regresijskih koeficientov so ocenjene natančno, vendar je priporočljivo imeti več kot 10 skupin. Velikost skupin ter velikost znotrajrazrednega koeficienta imata zanemarljiv vpliv na natančnost standardnih napak regresijskih koeficientov. V primeru ***komponent variance*** obstaja večja pristranskost kot pri regresijskih koeficientih. Enako kot v primeru regresijskih koeficientov znotrajrazredni koeficient nima vpliva na pristranskost, standardna napaka komponent variance se tudi tu povečuje z večanjem znotrajrazrednega korelacijskega koeficienta ter z zmanjševanjem števila skupin (N<20) ter velikosti skupin (n<20). Vpliv števila in velikosti skupin na natančnost standardnih napak komponent variance je nedvomno večji kot pri regresijskih koeficientih. Standardne napake varianc na makro nivoju so ocenjene kot premajhne tudi v primeru, ko imamo v vzorcu 50 skupin, v primeru zares majhnih vzorcev (npr. 10 skupin velikosti 5) so se izkazale za nesprejemljive. Velikost skupin ima manjši vpliv na stopnjo pokritosti kot

število skupin, natančnost se ne izboljša s povečanjem velikosti skupin. Razlika v znotrajrazrednem koeficientu nima vpliva na stopnjo pokritosti.

Osrednji del simulacijske študije predstavlja preverjanje statistične značilnosti testov ter natančnost ocen, kot jo pogojuje širina intervala. ***Moč statističnega testa regresijskih koeficientov*** sem računala s pomočjo dveh metod, ki sta se izkazali za zelo podobni, pri čemer je metoda standardne napake privedla do bolj konsistentnih rezultatov kot metoda nič/ena (predvsem v primeru manjših vzorcev), poleg tega zanjo potrebujemo manjše število simulacij. Za srednjo velikost regresijskega koeficienta spremenljivke na mikro nivoju (B=0.3) zadostno moč dosežemo že s 5 skupinami (z vsaj 30 enotami v skupini), če želimo doseči dovolj ozek interval zaupanja, potrebujemo vsaj 20 skupin (z vsaj 40 enotami v skupini). V primeru manjših koeficientov (B≤0.1) spremenljivke na makro nivoju ter interakcijske spremenljivke potrebujemo vsaj 50 skupin, z vsaj 80 enotami v skupini. Če želimo v primeru spremenljivke na makro nivoju doseči ožji interval zaupanja, potrebujemo vsaj 60 skupin. Za testiranje ***statistične moči variančnih parametrov*** sem uporabila t.i. test deviance oz. test razmerja verjetij. Če bi se želela prepričati, da bo varianca nagiba statistično značilna, bi potrebovala vsaj 30 skupin, z več kot 80 enotami v skupini. Opaznejše zožanje intervala zaupanja bi dosegla šele s 70 skupinami po 80 enot.

Izpeljava simulacij glede na vse kombinacije pogojev je zelo zamudna, zato se moramo o kombinacijah odločiti glede na proračun, ki nam je na voljo ter morebitne ostale omejujoče okoliščine. Na ta način simuliramo le omejen nabor vzorcev, ki se skladajo s temi omejitvami in preverimo obnašanje parametrov našega zanimanja (pristranskost, velikost standardnih napak, natančnost standardnih napak, moč testa, širina intervalov zaupanja). Najbolj optimalne kombinacije vzorcev lahko primerjamo tudi grafično in namesto zgolj opisnih statistik predstavimo celotne porazdelitve. Priporočam izvedbo analize občutljivosti, s katero lahko raziskovalec preveri posledice neuravnoteženosti vzorcev, manjšanje velikosti vpliva in podobno. V pričujoči raziskavi neuravnoteženost v smislu odklona od povprečnega števila skupin ni imela velikega vpliva na rezultate, kar pa ne velja za neuravnoteženost v smislu deleža majhnih skupin. Le-ta je privedla do širših intervalov zaupanja kot je to zaželeno v primeru spremenljivke na makro nivoju, v primeru interakcijske

spremenljivke pa do nezadostne moči statističnega testa. Ko načrtujemo primerno velikost vzorca, je ponavadi nerealistično pričakovati, da so vrednosti potrebnih populacijskih parametrov natančno znane. Priporočljivo je primerjati kvaliteto ocen tudi glede na različne vrednosti parametrov našega zanimanja. Spreminjanje velikosti regresijskih koeficientov nima vpliva na širino intervala, zelo močno pa je od velikosti vpliva odvisna moč testa.

Simulacijska študija nam je lahko v veliko pomoč pri raziskovanju učinkovitosti načrta, na ta način dobimo občutek za podatke, ki jih sploh še nismo zbrali. Za takšno simulacijo potrebujemo čas, pa tudi precej znanja statistike ter programiranja. Najtežja naloga za raziskovalca je nedvomno ta, da mora sklepati o parametrih, ki jih namerava vključiti v analizo, o njihovih vrednostih, porazdelitvah, poleg tega je pogosto omejen tudi s proračunom, znanjem in drugimi praktičnimi omejitvami. Za načrtovanje dobre večnivojske raziskave je priporočljivo najprej določiti primaren cilj raziskave, izraziti ta cilj v testiranem ali ocenjenem parametru in nato izbrati velikosti vzorcev, pri katerih bo ta parameter ocenjen z majhno standardno napako. Namesto lastne simulacijske študije lahko uporabimo programe, ki so na voljo za pomoč pri načrtovanju večnivojske analize, vendar mora raziskovalec tudi v tem primeru dobro poznati svoj teoretičen model in sklepati o parametrih, ki ga zanimajo, poleg tega pa mora obvladati tudi statistične podrobnosti same večnivojske analize.

Načrtovanje večnivojske raziskave tako predstavlja izredno kompleksen proces, ki zahteva sprejemanje in tehtanje številnih odločitev, ki v končni fazi pomembno vplivajo na kvaliteto in zanesljivost rezultatov. Če je ne izvedemo, lahko zapravimo ogromno časa in denarja v neuspelem poskusu, da ovržemo ničelno hipotezo, naše ocene pa lahko obdajajo intervali zaupanja, ki nas svojo širino 'spravljajo v zadrego'.

# 10 Literatura


Afshartous, David. 1995. *Determination of sample size for multilevel model design.* Perspectives on Statistics for Educational Research: Proceedings of the National Institute for Statistical Sciences (NISS), 1.oktober.

American Educational Research Association. 2006. *Standards for Reporting on Empirical Social Science Research in AERA Publications.* Washington, DC: Am. Educational Research Association.

American Psychological Association. 2001. *Publication Manual of the American Psychological Association.* Washington, DC: American Psychological Association.

Bainbridge, Timothy R. 1985. The Committee on Standards: precision and bias. *ASTM Standardization News 13:* 44-46.

Bassiri, Dina. 1988. *Large and small sample properties of maximum likelihood estimates for the hierarchical linear model.* Unpublished doctoral dissertation, Department of Counseling, Educational Psychology and Special Education, Michigan State University.

Brown Hendricks C. and Liao, Jason G. 1999. Principles for designing randomized preventive trials in mental health: an emerging developmental epidemiology paradigm. *American Journal of Community Psychology* 27 (9): 673-710.

Browne, William J. 1998. *Applying MCMC methods to multilevel models.* Unpublished doctoral dissertation, University of Bath, UK.

Browne, William J. in Draper, D. 2000. Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics* 15: 391–420.

Browne, William J., Golalizadeh Lahi M. in Parker, Richard. 2009. *A Guide to Sample Size Calculations for Random Effect Models via Simulation and the MLPowSim Software Package.*
Dostopno prek: *seis.bris.ac.uk/~frwjb/esrc/MLPOWSIMmanual.pdf* (9. september 2009).

Bryk, Anthony in Raudenbush, Stephen. 1992. *Hierarchical linear models.* Newbury Park, CA: Sage.

Busing, Frank. 1993. *Distribution characteristics of variance estimates in two-level models.* Unpublished manuscript, Leiden University, the Netherlands.

Casella, George in Lehmann, E. L. 1999. *Theory of Point Estimation*. Springer.

Cochran, Wiliam. 1977. *Sampling techniques*. New York: Wiley.

Cohen, Jacob. 1988. *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Cohen, Jacob. 1992. A power primer. *Psychological Bulletin* 133: 155-159.

Cohen, Jacob. 1994. The earth is round ($p < .05$). *American Psychologist* 49 (12): 997–1003.

Cohen, Michael. 1998. Determining sample sizes for surveys with data analyzed by hierarchical linear models. *Journal of Official Statistics* 14: 267-275.

Cools, Wilfred, Van den Noortgate, Wim in Onghena, Patrick. 2006. *Multi-Level Design Efficiency using simulation (ML-DEs).*
Dostopno prek: ppw.kuleuven.be/cmes/MLDEs.html (9. september 2009).

Cools, Wilfred, Van den Noortgate, Wim in Onghena, Patrick. 2008. ML-DEs: Multilevel design efficiency using simulation. *Behavior Research Methods* 40: 236-249.

Cools, Wilfred, Van den Noortgate, Wim in Onghena, Patrick. 2009. Design efficiency for imbalanced multilevel data. *Behavior Research Methods* 41: 192-203.

Debanne, Sara. 2000. The planning of clinical studies: bias and precision. *Gastrointestinal Endoscopy* 52: 821-822.

Dempster, Arthur P. in Tomberlin, Thomas J. 1980. The Analysis of Census under-count from a post-enumeration survey. V: Proceedings of the Conference on Census Undercount, str. 88-94.

Dempster, Arthur P., Laird, Nan M. Rubin in Donald B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* 39: 1-38.

Dempster, Arthur P., Rubin, Donald B. in Tsutakawa, Robert K. 1981. Estimation in Covariance Components Models. *Journal of the American Statistical Association* 76: 341-353.

Gelman, Andrew in Hill, Jennifer. 2007. *Data analysis using regression and multilevel/hierarchical models.* Cambridge University Press.

Goldstein, Harvey. 1986. Multilevel Mixed Linear Model Analysis Using Iterative Generalized Least Squares. *Biometrika* 73: 43-56.

Goldstein, Harvey. 1995/2003. *Multilevel statistical models*. London: Edwards Arnold. New York: Halstead.

Goodman, Steven in Berlin, James. 1994. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine* 121: 200–6.

Gulliford, Martin C., Ukoumunne, Obioha C. in Chinn, Susan. 1999. Components of variance and intraclass correlations for the design of community-based surveys and intervention studies. *American Journal of Epidemiology* 149: 876–883.

Hedeker, Donald, Gibbons, Robert in Waternaux, Christine. 1999. Sample size estimation for longitudinal designs with attrition: Comparing time-related contrasts between two groups. *Journal of Educational and Behavioral Statistics* 24: 70-93.

Hofmann David A. 1997. An overview of the logic and rationale of hierarchical linear models. *Journal of Management* 23: 723-744.

Howell, David. 2005. Power*. V: Encyclopedia of statistics in behavioral science, ur.* Everitt Brian S.*,* Howell David C.*,* 1558-1564*.* Chichester, U.K.*:* Wiley*.*

Hox, Joop. 1998. Multilevel modeling: when and why. V: *Classification, data analysis, and data highways*, ur. Balderjahn Ingo, in Schader, Martin, 147-154. New York: Springer-Verlag.

Hox, Joop. 2002. *Multilevel analysis: Techniques and applications.* Mahwah, NJ: Erlbaum.

Hunter, John E. in Schmidt, Frank L. 2004. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings.* Thousand Oaks, CA: Sage.

Kelley Ken in Maxwell Scott E. 2008. Sample size planning with applications to multiple regression: power and accuracy for omnibus and targeted effects. V *The Sage Handbook of Social Research Methods*, ur. Alasuutari Parti, Bickman Leonard, Brannen Julia, 166–92. London: Sage

Kelley Ken in Rausch Joseph R. 2006. Sample size planning for the standardized mean difference: accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods* 11 (4) :363–85.

Kelley Ken, Maxwell Scott E. in Rausch Joseph R. 2003. Obtaining power or obtaining precision: delineating methods of sample-size planning. *Evaluation* and the *Health Professions* 26 (3): 258–87.

Kelley, Ken in Maxwell, Scott. 2003. Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods* 8: 305-321.

Kim, Kate. S. 1990. *Multilevel data analysis: A comparison of analytical alternatives.* Unpublished doctoral dissertation, University of California at Los Angeles.

Kreft, Ita in de Leeuw, Jan. 1998. *Introducing multilevel modeling.* Sage: Thousand Oaks, CA.

Kreft, Ita, de Leeuw, Jan in Aiken, Leona S. 1995. The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research* 30: 1-21.

Longford, Nicholas T. 1987. A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika* 74: 817-827.

Longford, Nicholas T. 1993. *Random Coefficient Models.* Oxford, Clarendon Press.

Longford, Nicholas T. 1999. Standard errors in multilevel analysis. *Multilevel Modelling Newsletter* 11 (1): 10-13.

Maas, Cora in Hox, Joop. 2005. Sufficient sample sizes for multilevel modeling. *Methodology* 1: 86-92.

Maxwell Scott E., Ken Kelley in Joseph R. Rausch. 2008. Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation. *Annual Review of Psychology* 59: 537–563.

Maxwell, Scott E. 2004. The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychological Methods* 9 (2): 147–63.

Miller Rupert G. 1981. *Simultaneous Statistical Inference.* New York: Springer-Verlag.

Mok, Martin. 1995. Sample size requirements for 2-level designs in educational research. *Multilevel Modelling Newsletter* 7 (2): 11-15.

Muller, Keith E. in Benignus, Vernon A. 1992. Increasing Scientific Power with Statistical Power. *Neurotoxicology and Teratology* 14: 211–219.

Murphy, Kevin in Myors, Brett. 2003. *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests.* Mahwah, NJ: Erlbaum.

Murphy, Kevin in Myors, Brett. 2004. *Statistical Power Analysis.* London: Lawrence Erlbaum Associates, Publishers.

Muthén, Louis, in Muthén, Bengt. 2002. How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling* 9: 599-620.

R Development Core Team. 2004. *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.
Dostopno prek: http://www.R-project.org (19. april 2009).

Rasbash Jon, Browne William, Healy Michael, Cameron Bruce in Charlton Chris. 2005. *MLwiN (Version 2.02).*
Dostopno prek: http://www.cmm.bristol.ac.uk/MLwiN/ (19. april 2009).

Raudenbush, Stephen in Bryk, Anthony. 2002. *Hierarchical linear models: Applications and data analysis methods.* London: Sage Publications.

Raudenbush, Stephen in Liu, Xin. 2000. Statistical power and optimal design for multisite randomized trials. *Psychological Methods* 5: 199-213.

Raudenbush, Stephen in Liu, Xin. 2001. Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods* 6: 387-401.

Raudenbush, Stephen, Spybrook Jared, Liu Xin in Congdon Robert. 2005. *Optimal Design for Longitudinal and Multilevel Research.*
Dostopno prek: http://scholar.google.com (19. april 2009).

Raudenbush, Stephen. 1988. Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics* 13: 85-116.

Raudenbush, Stephen. 1997. Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods* 2: 173-185.

Reise, Steven P. in Duan, Naihua. 2003. Design issues in multilevel studies. V *Multilevel modeling: Methodological advances, issues, and applications*, ur. Reise S. P. in Duan N., 285-298. Mahwah, NJ: Lawrence Erlbaum.

Rose, Roderic A. in Bowen, Gary L. 2005. *Power for sample size in the evaluation of a whole-school educational change initiative.* Unpublished manuscript.
Dostopno prek: http://www.uncssp.org/rose_bowen_2005.pdf

Scherbaum, Charles A. in Ferreter, Jane M. 2008. Estimating statistical power and sample size requirement for organizational research using hierarchical linear models. *Organizational Research Methods* 11 (4): 659 – 681.

Sedlmeier, Peter in Gigerenzer, Gerd. 1989. Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin 105:* 309-316.

Self Steven G. in Liang Kung-Yee (1987). Large sample properties of the maximum likelihood estimator and the likelihood ratio test on the boundary of the parameter space. *Journal of the American Statistical Association* 82: 605-611.

Snijders, Tom in Bosker, Roel. 1993. Standard errors and sample sizes for two-level research. *Journal of Educational Statistics* 18: 237-259.

Snijders, Tom in Bosker, Roel. 1999. *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London: Sage.

Snijders, Tom. 2005. Power and sample size in multilevel linear models. V: *Encyclopedia of Statistics in Behavioral Science*, ur. Everitt Brian S. in Howell David C.*,* 1570-1573. Chicester: Wiley.

Van den Noortgate, Wim in Onghena, Patrick. 2003. Multilevel meta-analysis: A comparison with traditional meta-analytical procedures. *Educational and Psychological Measurement* 63: 765-790.

Van den Noortgate, Wim in Onghena, Patrick. 2006. Analysing repeated measures data in cognitive research: A comment on regression coefficient analyses. *European Journal of Cognitive Psychology* 18: 937-952.

Van der Leeden, Rien in Busing, Frank. 1994. *First iteration versus IGLS RIGLS estimates in two-level models: A Monte Carlostudy with ML3*. Unpublished manuscript, Leiden University, the Netherlands.

Van der Leeden, Rien, Busing, Frank in Meijer, Erik (1997, April). *Applications of bootstrap methods for two-level models*. Paper presented at the Multilevel Conference, Amsterdam.

Wilkinson, Leland. 1999. Statistical methods in psychology journals: guidelines and explanations. *American Psychologist* 54 (8): 594–604.