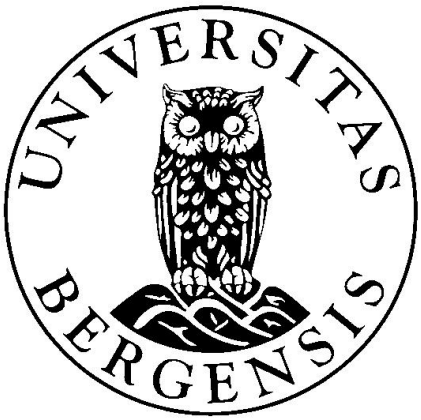
Testmetoder for identifisering av publikasjonsbias i metaanalyser

Master i matematisk statistikk

Miriam Gjerdevik



Universitetet i Bergen

Matematisk institutt

24. oktober 2012

Takk!

Først og fremst fortjener veilederen min, Ivar Heuch, en stor takk. Han er dyktig, tålmodig, hjelpsom og motiverende! Ikke minst vil jeg takke Heuch for at jeg fikk delta på ISCB 33. Dette var både gøy og lærerikt!

Jeg ønsker også å takke Henning Lohne for kombinatorikkhjelp i anledning beviset i Seksjon 2.5.

Mange medstudenter har bidratt til en uforglemmelig tid som student. Jeg vil særlig takke Leiv Magne, Sindre, Jon, Erik, Mirjam og Ingrid, som har vært gode støttespillere både faglig og sosialt.

Innhold

Innledning **1**

1 Metaanalyse, publikasjonsbias og funnelplott **3**

1.1 Metaanalyse . . . 3

1.2 Publikasjonsbias . . . 4

1.3 Funnelplott . . . 5

2 Kendalls tau **7**

2.1 Felles behandling av Pearsons korrelasjon, Kendalls tau og Spearmans rho . . . 7

2.1.1 Pearsons produktmomentkorrelasjon . . . 8

2.1.2 Kendalls tau . . . 8

2.1.3 Spearmans rho . . . 9

2.2 Valg av test . . . 10

2.3 Utledning av den asymptotiske fordelingen til Kendalls tau under nullhypotesen 11

2.4 Enda en utledning av variansen til Kendalls tau under nullhypotesen . . . 16

2.5 Forventningen til $a_{ij}a_{ik}$ . . . 19

3 Beskrivelse av Begg og Mazumdars testmetode og simuleringssituasjoner **21**

3.1 Motivasjon for testmetoden . . . 21

3.2 Testmetoden til Begg og Mazumdar . . . 21

3.3 Faktorer som kan påvirke testens styrke . . . 22

3.4 Simuleringer . . . 22

3.4.1 Seleksjonsmodeller . . . 23

3.4.2 Parametervalg . . . 24

3.4.3 Utføring . . . 24

3.5 Hypoteser forbundet med testen for publikasjonsbias . . . 26

4 Simuleringsresultater for Begg og Mazumdars testmetode **27**

4.1 Simuleringsresultater for metaanalyser med publikasjonsbias . . . 27

4.1.1 Kontroll av Begg og Mazumdars simuleringsresultater . . . 27

4.1.2 Vurdering av testmetodens egenskaper . . . 32

4.2 Simuleringsresultater for metaanalyser uten publikasjonsbias . . . 33

4.3 Reell variansfordeling . . . 35

5 Mulige årsaker til det feilaktige signifikansnivået for testmetoden introdusert av Begg og Mazumdar **37**

5.1 Signifikansnivået til rangkorrelasjonstesten basert på Kendalls tau når alle forutsetninger er oppfylt . . . 37
5.2 Sannsynlighetsfordelingen til Kendalls tau i Begg og Mazumdars testmodell . . 39
5.3 Spearmans rho . . . 40
5.4 Antakelser . . . 41
5.5 Misbruk av notasjon . . . 41
5.6 Variansen til $t_i - \bar{t}$ gitt $v_1, v_2, \ldots, v_k$ . . . 41
5.7 Fordelingen til t_i^* gitt $v_1, v_2, \ldots, v_k$ i Begg og Mazumdars testsituasjon . . . 43
5.8 Er t_i^* uavhengig av $v_1, v_2, \ldots, v_k$ under nullhypotesen i testsituasjonen til Begg og Mazumdar? . . . 44
5.9 Bivariat fordeling . . . 44
5.10 Uavhengige par . . . 45
5.11 Kan $\mathbf{t}^*$ og $\mathbf{v}$ antas uavhengige under nullhypotesen? . . . 48

6 Hvordan påvirkes signifikansnivået til rangkorrelasjonstesten basert på Kendalls tau ved brudd på de ulike forutsetningene? **51**

6.1 Er nivået ukorrekt fordi Begg og Mazumdar formelt sett lar variansene være faste størrelser? . . . 51
6.2 Hvordan påvirkes nivået av at de forskjellige observasjonsparene ikke er uavhengige? . . . 53
6.3 Hvordan påvirkes nivået dersom t_i^* og v_i ikke er uavhengige under nullhypotesen? 55

7 Forslag til mulig forbedring av Begg og Mazumdars testmetode **57**

7.1 Kort om hypotesetesting og evaluering av ulike testmetoder . . . 57
7.2 Motivasjon og forslag til mulig forbedring av testmetode . . . 60
7.3 Forutsetninger som brytes i den ustandardiserte testen under nullhypotesen . . 63
7.4 Simuleringsresultater og vurdering av den ustandardiserte testprosedyren 64

8 Forslag til forbedring av testmetoder basert på den simulerte fordelingen til Kendalls tau **73**

8.1 Beskrivelse av testmetoder med tilnærmet korrekt nivå . . . 73
8.2 Utfordringer knyttet til de korrigerte testmetodene . . . 76
8.3 Simuleringer, resultater og sammenlikning av testobservatorenes informasjon . . 76
8.3.1 Er den simulerte fordelingen til Kendalls tau robust dersom variansene systematisk underestimeres? . . . 77

8.3.2 Sammenlikning av informasjonen til den standardiserte og ustandardiserte testobservatoren . . . 78

9 Regresjon 83

9.1 Enkel lineær regresjon . . . 83
9.1.1 Minste kvadraters metode . . . 83
9.1.2 Utledning av testobservatorer . . . 84
9.1.3 Konsekvenser ved brudd på forutsetninger . . . 88
9.2 Vektet lineær regresjon . . . 95
9.2.1 Vektet minste kvadraters metode og utledning av testobservatorer . . . 95
9.2.2 Faremomenter ved bruk av vektet regresjon . . . 99

10 Regresjonsbaserte tester for å identifisere publikasjonsbias 101

10.1 Beskrivelse av metaanalysesituasjon . . . 101
10.2 Eggers regresjonsmetoder . . . 102
10.2.1 Metoder . . . 102
10.2.2 Motivasjon . . . 102
10.2.3 Utfordringer knyttet til Eggers regresjonsmetoder . . . 105
10.3 Funnelplottregresjon . . . 106
10.3.1 Tidligere introduserte metoder basert på funnelplottregresjon . . . 106
10.3.2 Ny metode basert på funnelplottregresjon . . . 107

11 Ny simuleringssituasjon, simuleringsresultater og diskusjon 109

11.1 Simuleringer . . . 109
11.1.1 Seleksjonsmodeller . . . 109
11.1.2 Parametervalg . . . 109
11.1.3 Utføring . . . 110
11.2 Simuleringsresultater . . . 112
11.2.1 Simuleringsresultater for metaanalyser uten publikasjonsbias, konfigurasjon A og B . . . 112
11.2.2 Simuleringsresultater for metaanalyser med publikasjonsbias, konfigurasjon A og B . . . 114
11.2.3 Effekten av å øke antall studier per metaanalyse, konfigurasjon C og D . 114
11.2.4 Effekten av å øke studienes sampelstørrelse, konfigurasjon E og F . . . 115
11.3 Asymmetri . . . 127
11.4 Testmetodene basert på rangkorrelasjon . . . 129
11.5 Anbefaling av testmetode . . . 130

11.6 En enkel sammenlikning av Eggers regresjonsmetoder og de korrigerte rangkorrelasjonstestene i Begg og Mazumdars simuleringssituasjonen 132

12 Ortogonal regresjon som mulig forbedring til Eggers uvektede regresjonsmetode 137

12.1 Ortogonal minste kvadraters metode 138
12.2 Testobservator for inferens om skjæringspunktet, samt dens fordeling 139
12.3 Kommentarer til bruk av ortogonal regresjon ved testing for publikasjonsbias . 141
12.4 Simuleringsresultater og vurdering av testmetoden 142

13 Oppsummering og videre arbeid 145

A Kort om konfidensintervaller 147

B Odds-ratio 151

C Eksempel på simuleringskode brukt for å kontrollere Begg og Mazumdars simuleringsresultater 153

D Eksempel på simuleringskode som viser hvordan nivået til rangkorrelasjonsmetodene kan tilpasses 157

E Eksempel på simuleringskode brukt for å kontrollere simuleringsresultatene til Macaskill et al. 159

F Akseptert sammendrag i anledning ISCB 33 165

Referanser 167

Innledning

Begg og Mazumdars artikkel [8] danner utgangspunktet for denne oppgaven. Artikkelen omhandler tematikken publikasjonsbias i metaanalyser. Forfatterne introduserer en statistisk testmetode for identifisering av publikasjonsbias i metaanalyser. Testen introduseres på et intuitivt grunnlag. En forventer at skjevhet i funnelplottet impliserer publikasjonsbias. Asymmetri i funnelplottet undersøkes ved hjelp av en rangkorrelasjonstest basert på Kendalls tau. Begg og Mazumdar utforsker testmetodens egenskaper i nokså generelle testsituasjoner. Situasjonene avviker på enkelte områder fra en reell metaanalysesammenheng.

Opprinnelig var målet med masteroppgaven først og fremst å forstå og utdype artikkelen skrevet av Begg og Mazumdar [8]. Deretter skulle jeg arbeide videre med problemstillinger introdusert i artikkelen. Et naturlig utgangspunkt var derfor å kontrollere Begg og Mazumdars simuleringsresultater. Simuleringsresultatene viser at testens signifikansnivå ikke er korrekt. Dette er en utfordring Begg og Mazumdar ikke framhever uttrykkelig i sin artikkel. Problemstillingen forandrer med dette karakter. Jeg har ikke funnet litteratur som utdyper problemet med testmetodens nivå og ønsker derfor å undersøke mulige årsaker til det feilaktige signifikansnivået. Videre ønsker jeg å korrigere nivået og å forbedre den opprinnelige testen. Modifiserte testmetoder basert på rangkorrelasjon introduseres.

Det er nyttig å undersøke egenskapene til rangkorrelasjontestene i andre situasjoner enn dem Begg og Mazumdar tar for seg. Av den grunn er det naturlig også å fordype seg i artikkelen skrevet av Macaskill et al. [35]. Her har vi en konkret statistisk modell, hvor 2×2-tabeller simuleres. Ulike regresjonsbaserte testmetoder presenteres. Testmetodene har alle som mål å teste for publikasjonsbias i metaanalyser. Disse testene er introdusert av Egger et al. [15] og Macaskill et al. [35]. Jeg modifiserer noen av disse metodene i et forsøk på å forbedre testenes egenskaper. De ulike testmetodene presentert i oppgaven vurderes og sammenliknes.

Kapittel 1 gir en kort introduksjon om metaanalyser, publikasjonsbias og funnelplott. Dette er bakgrunnstoff som er nødvendig for å sette seg inn i oppgavens tematikk.

I Kapittel 2 får leseren kjennskap til den generelle rangkorrelasjonstesten basert på Kendalls tau. Kunnskap om denne testen gjør det enklere å forstå problemene som senere dukker opp når rangkorrelasjon anvendes som en teknikk for å avdekke publikasjonsbias. Jeg utleder den asymptotiske fordelingen til Kendalls tau under nullhypotesen om ingen publikasjonsbias. Deler av bevisene er behandlet annerledes enn i den litteraturen jeg har kjennskap til.

Kapittel 3 omhandler artikkelen til Begg og Mazumdar [8]. Testmetoden for å identifisere publikasjonsbias i metaanalyser introduseres. Videre beskrives simuleringsprosedyren forfatterne nytter for å undersøke testens egenskaper.

Simuleringsresultatene til Begg og Mazumdar [8] kontrolleres i Kapittel 4.

I Kapittel 5 undersøkes mulige årsaker til det feilaktige signifikansnivået. Flere forutsetninger for å utføre en rangkorrelasjonstest brytes.

Hvordan signifikansnivået påvirkes av brudd på de ulike forutsetningene utforskes nærmere i Kapittel 6.

Signifikansnivået til Begg og Mazumdars test for publikasjonsbias samsvarer ikke med det nominelle. Konsekvensene kan være alvorlige og kan ikke neglisjeres. Argumentasjonen for dette gis i Kapittel 7. På bakgrunn av denne argumentasjonen foreslår jeg en mulig forbedring til forfatternes testmetode.

I Kapittel 8 korrigeres signifikansnivået til Begg og Mazumdars testmetode. Jeg nytter den simulerte fordelingen til Kendalls tau, betinget på de estimerte variansene. Dette kapittelet avslutter mer eller mindre fortellingen om rangkorrelasjonstester basert på Kendalls tau som metoder for å identifisere publikasjonsbias i metaanalyser. Jeg returnerer likevel til disse testene i Kapittel 11.

Kapittel 9 gir grunnleggende teori om lineær regresjonsanalyse. Teorien danner grunnlaget for å undersøke regresjonsbaserte testmetoder for å avdekke publikasjonsbias.

Kapittel 10 introduserer testmetoder basert på regresjon. Fordeler og ulemper ved de ulike testmetodene diskuteres kort.

Egenskapene til de ulike testmetodene introdusert gjennom oppgaven utforskes ved hjelp av simuleringer i Kapittel 11. En ny simuleringssituasjon presenteres. Testobservatorene vurderes og sammenliknes.

I Kapittel 12 undersøker jeg om ortogonal regresjon kan nyttes for å forbedre Eggers uvektede regresjonsmetode.

I Kapittel 13 vil jeg oppsummere og kort introdusere enkelte retninger for videre arbeid.

I oppgaven beregnes konfidensintervaller for ulike parametre. Vedlegg A omhandler grunnleggende teori om konfidensintervaller og beregning av disse. Vedlegg B definerer odds-ratio og utleder den asymptotiske fordelingen til log-odds-ratio-estimatoren. Eksempel på simuleringskode brukt for å kontrollere Begg og Mazumdars resultater finnes i Vedlegg C. Vedlegg D viser kode som eksemplifiserer hvordan algoritmen for å korrigere nivået til rangkorrelasjonstestene kan implementeres i praksis. Vedlegg E gir eksempel på kode brukt for å kontrollere resultatene til Macaskill et al.

I august 2012 holdt jeg foredrag C.34.1 [25] på ISCB 33. Foredraget omhandlet utfordringene knyttet til testmetoden introdusert av Begg og Mazumdar og hvordan feilratene kan forbedres. Det aksepterte sammendraget inkluderes i Vedlegg F.

1 Metaanalyse, publikasjonsbias og funnelplott

Metaanalyse, publikasjonsbias og funnelplott er tre begreper som er sentrale for å forstå oppgavens tematikk. Jeg ønsker å gi leseren en kort innføring i disse begrepene, men tar ikke sikte på å forklare dem inngående. Hvordan en kan utføre metaanalyser, forskjellige metaanalysemodeller, fordeler og ulemper vil ikke bli gjennomgått. Dette har ikke betydning for videre lesing.

1.1 Metaanalyse

På verdensbasis utføres mange nærmest identisk like studier. Ett eksempel er studier som ser på sammenhengen mellom kaffedrikking og lungekreft [50]. En ønsker ofte å sammenfatte resultatene i de forskjellige studiene til ett felles resultat. Dette kan gjøres ved hjelp av en metaanalyse.

En metaanalyse er en systematisk metode for å evaluere statistiske data basert på resultater fra flere uavhengige studier som behandler det samme problemet [1]. En metaanalyse kan defineres som en samling av statistiske teknikker for å oppsummere enkeltresultater fra flere rapporter innenfor ett område. Statistiske teknikker gjør det mulig å tillegge data fra enkeltundersøkelser ulik vekt. Studier med stor sampelstørrelse vil ofte vektes mer enn studier med liten sampelstørrelse [2]. I en metaanalyse kvantifiseres relevante resultater fra hver studie på en slik måte at de resulterende verdiene kan aggregeres og sammenliknes [54]. Forenklet kan en gjerne definere en metaanalyse som en analyse av analyser. Medisinske tidsskrifter er kjente publiseringssteder for metaanalyser.

Resultater fra en metaanalyse er basert på flere studier. En kan derfor konkludere med høyere grad av sikkerhet. Likevel vil ikke en metaanalyse bestående av flere små studier kunne predikere resultatene til en stor studie.

Resultater fra ulike studier vil ofte sprike. Det kan være forskjellige årsaker til dette. Noen studier er muligens for små. Ulike pasientgrupper kan ha deltatt i studiene. Vi tenker vi har uavhengige studier hvor alle forsøker å estimere sammenhengen mellom kaffedrikking og lungekreft. Røyking kan medføre økt risiko for lungekreft. Hva om noen studier inkluderer pasienter hvor flere røyker, mens andre kun lar ikke-røykere delta?

Aldersforskjeller, kjønnsforskjeller og ulikheter i gjennomføring av en studie er eksempler på det en kaller klinisk eller metodologisk heterogenitet. Slike forskjeller kan føre til uoverensstemmelser mellom resultatene til ulike studier, men trenger ikke være kilden til disse ulikhetene. Statistisk heterogenitet eksisterer når den sanne effekten som blir evaluert er ulik i forskjellige studier. Dette kan i enkelte tilfeller oppdages dersom variasjonen mellom resultatene er større enn hva en kan forvente ved tilfeldighet. Omfanget av statistisk heterogenitet

i en metaanalyse kan gjøre det vanskelig å trekke generelle konklusjoner [21].

I min oppgave har metaanalysene en modell med faste effekter. Her antar vi at den underliggende effekten er den samme for alle studier. Dette står i kontrast til en modell med tilfeldige effekter, hvor den underliggende effekten kan variere mellom studier. Leseren henvises til artikkelen av Dersimonian og Laird [11] for mer informasjon om modellen med tilfeldige effekter.

En modell med faste effekter forsøker å svare på hvor stor den gjennomsnittlige sanne effekten er i en metaanalyse bestående av k studier. Et vektet gjennomsnitt av estimatene fra de ulike studiene kan brukes til å estimere den sanne effekten. Vektene er gjerne inversen til sampelvariansen til de observerte effektestimatene [54]. I tilfeller hvor vi har 2×2-tabeller, kan effekten også estimeres ved hjelp av Mantel-Haenszels estimator. Denne estimatoren defineres i Seksjon 10.1.

Metaanalyser er kanskje spesielt utsatt for kritikk grunnet publikasjonsbias. Hoveddelen av oppgaven behandler nettopp dette temaet. Publikasjonsbias er derfor berettiget sin egen seksjon.

1.2 Publikasjonsbias

Bias kan oversettes til skjevhet. Skjevhet kan være et stort problem i alle ledd av en forskningsprosess. Hovedsaklig kan skjevhet i forskning føre til resultater som ikke samsvarer med virkeligheten. Begg og Berlin [7] dokumenterer at publikasjonsbias er et reelt problem. Det finnes ulike former for bias. Intervjuerbias, frafallsbias og hukommelsesbias er noen få.

Publikasjonsbias forekommer når de publiserte studiene som inngår i en metaanalyse ikke representerer alle studiene om det aktuelle temaet. Språkbias kan være en årsak til publikasjonsbias. En engelsktalende forsker kan ha problemer med å finne aktuelle studier publisert på finsk eller kinesisk. Selv om dette kan være problematisk, er risikoen for feilaktige resultater trolig større når årsaken til publikasjonsbias er at studier som støtter opp om en nullhypotese sjeldnere blir publisert enn studier som går i favør av en alternativ hypotese.

I praksis er sannsynligheten for at en studie publiseres assosiert med dens resultater. Studier med liten sampelstørrelse og lav statistisk presisjon blir sjeldnere publisert enn studier med stor sampelstørrelse og høy statistisk presisjon [35]. Det samme gjelder studier som ikke viser effekt. Begg [5] beskriver et scenario hvor et antall forskere uavhengig gjennomfører identiske studier for å estimere en effekt. Effektestimatene vil variere grunnet tilfeldig variasjon, statistisk heterogenitet eksisterer ikke. Forskeren som gjennomførte studien som viser den mest signifikante effekten, vil være den som mest sannsynlig publiserer resultatene. Dette kan gi skjevhet i estimatet for den underliggende effekten. Skjevhetens størrelse er assosiert med studiens sampelstørrelse. Vi bør være ekstra bekymret for publikasjonsbias i metaanalyser som

inneholder mange små studier. Publikasjonsbias påvirker store individuelle studier i mindre grad.

Et faremoment med publikasjonsbias, i tillegg til et biased estimat for den underliggende effekten, er at den aggregerte sampelstørrelsen kan være stor. Resultatene i en metaanalyse er tilsynelatende nøyaktige og presise, men er ikke mindre biased av den grunn [7].

Det er nyttig å søke etter relevante studier som ikke er publiserte. Dette kan begrense utbredelsen av publikasjonsbias, men er dessverre ofte en vanskelig oppgave. Det er behov for metoder som kan avdekke publikasjonsbias på bakgrunn av dataene i de tilgjengelige studiene [35].

1.3 Funnelplott

Funnelplott er ofte brukt for å vurdere risikoen for publikasjonsbias. Funnelplott er en grafisk figur som viser et mål for sampelstørrelsen til de ulike studiene i en metaanalyse plottet mot de estimerte effektstørrelsene. Hva menes med et mål for studienes sampelstørrelse? Egger et al. [15] nytter presisjon. Presisjonen defineres ved $1/\sqrt{v_i}$, hvor v_i er variansen til studie i. Noen bruker studiens sampelstørrelse direkte, mens andre liker inversen til behandlingseffektens varians. Effektstørrelse brukes her som en fellesbetegnelse for resultatet i en metaanalyse. Effektstørrelsen kan blant annet være målt i odds-ratio, relativ risiko eller hazard-ratio. Odds-ratio behandles nærmere i Vedlegg B. En behandlingseffekt kan eksempelvis være målt senkning av blodtrykk i kliniske forsøk eller log-odds-ratio i epidemiologi.

I denne oppgaven antar en at alle studiene i metaanalysen estimerer den samme effekten. De estimerte effektstørrelsene bør være fordelt rundt den sanne verdien av effekten om effektestimatene er representert på en passende skala. Hvis effektestimatene er målt i odds ratio, må vi bruke en logaritmisk skala for at denne symmetrien skal vise seg.

Kort fortalt er funnelplottet basert på det faktum at presisjonen ved estimering av den underliggende effekten vil øke når sampelstørrelsen til hver enkelt studie øker [15]. Estimatenes presisjon er høyere for studier med stor sampelstørrelse enn for studier med liten sampelstørrelse [31]. Dersom vi har studienes sampelstørrelse langs den vertikale aksen, bør små studier langs bunnen av figuren ligge spredt rundt den ukjente, sanne effekten. Studiene bør ligge nærmere den sanne verdien jo større studiene blir [8]. I en metaanalyse vil det naturlig være mange små studier og færre store studier [35]. Uten publikasjonsbias eller heterogenitet forventer en at plottet likner en omvendt symmetrisk trakt [15].

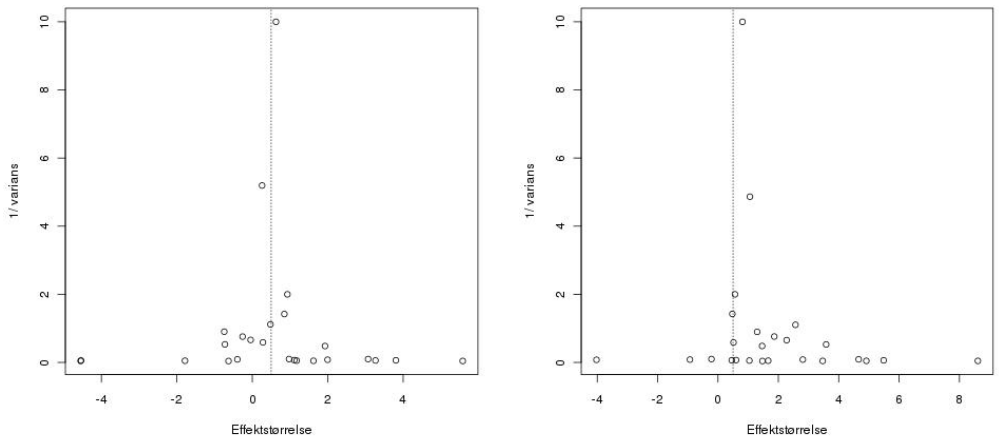
Det er nærliggende å anta at den samme symmetrien gjelder dersom vi har et plott med den ekte variansen eller den virkelige presisjonen langs den vertikale aksen. Formen vil likne en symmetrisk trakt dersom variansen avsettes langs den vertikale aksen. Studier med høy presisjon og liten varians vil estimere den sanne effekten med større grad av nøyaktighet enn

studier med lav presisjon og stor varians.

Dersom studier med liten sampelstørrelse og lav presisjon med ingen eller negativ effekt har mindre sannsynlighet for å bli publisert, vil grafen være skjev. Enkelte små studier med effekt nær null vil mangle i metaanalysen [35]. En studies sampelstørrelse, presisjon eller varians er da korrelert med effektstørrelsen. Funnelplottet er ikke symmetrisk. Asymmetriske funnelplott impliserer publikasjonsbias.

Funnelplottet har tidligere blitt brukt som en uformell metode for å identifisere publikasjonsbias i metaanalyser. En har undersøkt skjevhet og asymmetri visuelt. Denne visuelle inspeksjonen er subjektiv. Statistiske tester er innført for å kunne avdekke eventuell asymmetri i funnelplottet og dermed oppdage publikasjonsbias ved hjelp av mer objektive metoder.

Tang et al. [50] nytter funnelplott for å identifisere publikasjonsbias i en metaanalyse hvor sammenhengen mellom kaffedrikking og lungekreft undersøkes. Visuell inspeksjon viser ikke antydning til publikasjonsbias. Begg [5] refererer til en metaanalyse av Raudenbush [44]. Metaanalysen estimerer effekten av lærers forventning på elevs IQ. Begg konstruerer et funnelpott på bakgrunn av dataene gitt av Raudenbush. Visuell inspeksjon av funnelplottet gir grunn til å mistenke at det eksisterer små studier med liten effekt som ikke er blitt publisert. Vandenbroucke [52] utfordrer en tidligere publisert artikkel som vurderer sammenhengen mellom passiv røyking og risiko for lungekreft. Også her viser funnelplottet antydning til publikasjonsbias. Figur 1 viser typiske funnelplott med og uten publikasjonsbias, her med inversen til variansen langs den vertikale aksen.



(a) Funnelplott uten publikasjonsbias. (b) Funnelplott med publikasjonsbias.

Figur 1: Eksempel på funnelplott med og uten publikasjonsbias basert på simulerte metaanalyser. Den vertikalt stiplede linjen indikerer den sanne behandlingseffekten.

2 Kendalls tau

Ofte har en behov for å måle graden av sammenheng mellom to eller flere variable. Fra og med første grunnkurs i statistikk har jeg brukt kovarians og korrelasjon til dette formålet. Disse måler graden av lineær sammenheng mellom to stokastiske variable X og Y. Korrelasjonen mellom X og Y kalles Pearsons produktmomentkorrelasjon eller bare Pearsons korrelasjon.

Sammenhengen mellom to variable kan også måles ved hjelp av Kendalls rangkorrelasjonskoeffisient eller Spearmans rangkorrelasjonskoeffisient. Jeg kaller dem Kendalls tau og Spearmans rho henholdsvis. Disse måler graden av monoton sammenheng mellom X og Y. Det er en positiv monoton sammenheng dersom en økende verdi hos en variabel alltid assosieres med en økende verdi hos den andre variabelen. På samme måte er det en negativ monoton sammenheng dersom en økende verdi hos den ene variabelen alltid er assosiert med en minkende verdi hos den andre variabelen [48]. Begg og Mazumdar [8] konstruerer en test for å identifisere publikasjonsbias i metaanalyser ved å undersøke om det er en monoton sammenheng mellom effektestimatene og deres varianser. Forfatterne nytter Kendalls tau.

Målet med dette kapittelet er ikke å gjennomgå all teori som finnes om Kendalls tau. Jeg vil først og fremst ta for meg teorien som er sentral i forhold til Begg og Mazumdars artikkel [8]. I tillegg inngår noe teori om Pearsons korrelasjon og Spearmans rho. Jeg ønsker å sette Kendalls tau inn i en helhetlig ramme.

2.1 Felles behandling av Pearsons korrelasjon, Kendalls tau og Spearmans rho

Selv om Pearsons korrelasjon, Spearmans rho og Kendalls tau er ulike mål for sammenhengen mellom to variable, har koeffisientene mange fellestrekk. Jeg vil utlede dem fra en generell korrelasjonskoeffisient. Jeg tar utgangspunkt i utledningene gitt av Kendall og Gibbons [27], men velger å gå noe mer i detalj.

La $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ være de observerte verdiene av de todimensjonale stokastiske vektorene $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. De n stokastiske parene er uavhengige. Hvert par har den samme kontinuerlige fordelingen. For hvert observasjonspar, x_i og x_j, tildeler vi en X-score som vi betegner a_{ij}. Vi tildeler også en Y-score til observasjonsparene bestående av y_i og y_j. Denne betegnes b_{ij}. Vi definerer $a_{ij} = -a_{ji}$, $b_{ij} = -b_{ji}$ og $a_{ii} = b_{ii} = 0$.

Den generaliserte korrelasjonskoeffisienten Γ defineres ved

$$\Gamma = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}b_{ij}}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2 \sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}^2}}. \tag{1}$$

Den teoretiske sannsynligheten for at $x_i = x_j$ eller $y_i = y_j$ er lik null for $i \neq j$. Dette følger fordi X_i, X_j, Y_i og Y_j antas å ha en kontinuerlig fordeling. I praktiske situasjoner risikerer man at to eller flere observasjoner er så like at vi ikke klarer å skille dem fra hverandre. Dette kalles ties. Utfordringer knyttet til ties vil ikke diskuteres her.

2.1.1 Pearsons produktmomentkorrelasjon

Jeg starter med å utlede Pearsons korrelasjonskoeffisient. I dette tilfellet defineres $a_{ij} = x_j - x_i$ og $b_{ij} = y_j - y_i$. Jeg setter disse uttrykkene inn i den generelle korrelasjonsformelen (1) og får

$$\begin{aligned}
\Gamma = r &= \frac{\sum\sum (x_j - x_i)(y_j - y_i)}{\sqrt{\sum\sum (x_j - x_i)^2 \sum\sum (y_j - y_i)^2}} \\
&= \frac{\sum\sum x_i y_i + \sum\sum x_j y_j - \sum\sum (x_i y_j + x_j y_i)}{\sqrt{2n\sum x_i^2 - 2(\sum x_i)^2}\sqrt{2n\sum y_i^2 - 2(\sum y_i)^2}} \\
&= \frac{2n(\sum x_i y_i) - 2\sum\sum x_i y_j}{\sqrt{2n\sum x_i^2 - 2(\sum x_i)^2}\sqrt{2n\sum y_i^2 - 2(\sum y_i)^2}} \\
&= \frac{n(\sum x_i y_i) - \sum x_i \sum y_j}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}.
\end{aligned}$$

Som tidligere nevnt er de n parene $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ uavhengige og identisk fordelte stokastiske vektorer. Jeg antar at $\mathrm{E}(X_i^2)$ og $\mathrm{E}(Y_i^2)$ eksisterer for $i = 1, 2, \ldots, n$. Da eksisterer også $\mathrm{E}(X_i), \mathrm{E}(Y_i)$ og $\mathrm{E}(X_i Y_i)$ [37].

Khinchins setning, også kalt De store talls lov, gir at $\bar{X} \xrightarrow{p} \mathrm{E}(X)$, $\bar{Y} \xrightarrow{p} \mathrm{E}(Y)$, $\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{p} \mathrm{E}(X^2)$, $\frac{1}{n}\sum_{i=1}^{n} Y_i^2 \xrightarrow{p} \mathrm{E}(Y^2)$ og $\frac{1}{n}\sum_{i=1}^{n} X_i Y_i \xrightarrow{p} \mathrm{E}(XY)$. Symbolet $\xrightarrow{p}$ betegner konvergens i sannsynlighet. Det følger at

$$\Gamma \xrightarrow{p} \rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

2.1.2 Kendalls tau

Det er nødvendig med flere definisjoner for å utlede Kendalls tau, t, fra den generelle korrelasjonskoeffisienten. Vi observerer $x_1, x_2, \ldots, x_n$. Rangen til observasjon i og j betegnes p_i og p_j henholdsvis. Vi antar at vi ikke har ties. Det følger at $p_i \neq p_j$. Vi definerer $a_{ij} = 1$ dersom

$p_j > p_i$, og $a_{ij} = -1$ hvis $p_j < p_i$. For observasjonene $y_1, y_2, \ldots, y_n$, defineres størrelsene tilsvarende. Rangen til observasjon i og j betegnes henholdsvis q_i og q_j. Vi setter $b_{ij} = 1$ dersom $q_j > q_i$, og $b_{ij} = -1$ dersom $q_j < q_i$. Vi antar $q_i \neq q_j$. Se Seksjon 2.4 for en mer oversiktlig definisjon av a_{ij} og b_{ij}. Videre kaller vi (x_i, y_i) og (x_j, y_j) konkordante dersom $(x_i - x_j)(y_i - y_j) > 0$. Dette tilsvarer at $x_i < x_j$ når $y_i < y_j$, eller at $x_i > x_j$ når $y_i > y_j$. Parene er diskordante dersom $(x_i - x_j)(y_i - y_j) < 0$, altså når $x_i < x_j$ samtidig som $y_i > y_j$, eller $x_i > x_j$ samtidig som $y_i < y_j$.

Telleren i det generelle uttrykket (1) kan uttrykkes ved $\sum\sum a_{ij}b_{ij} = 2(C-D) = 2S$, hvor $S = C - D$, C er antall konkordante par og D er antall diskordante par. Fordi $a_{ij}^2 = b_{ij}^2 = 1$ for $i \neq j$, følger det at $\sum\sum a_{ij}^2 = \sum\sum b_{ij}^2 = n(n-1)$. Innsetting i den generelle formelen gir

$$\Gamma = t = \frac{2S}{n(n-1)}.$$

Kendalls tau, t, er sannsynligheten for konkordans minus sannsynligheten for diskordans for et par av observasjoner, (x_i, y_i) og (x_j, y_j), trukket tilfeldig fra utvalget [38].

La $(X_1, Y_1), (X_2, Y_2)$ være uavhengige og identisk fordelte stokastiske vektorer. Nelsen [38] og Kruskal [28] definerer parameterversjonen til Kendalls tau som sannsynligheten for konkordans minus sannsynligheten for diskordans:

$$\tau = P\big((X_1 - X_2)(Y_1 - Y_2) > 0\big) - P\big((X_1 - X_2)(Y_1 - Y_2) < 0\big). \tag{2}$$

2.1.3 Spearmans rho

Når en skal utlede Spearmans rho, lar en $a_{ij} = p_j - p_i$ og $b_{ij} = q_j - q_i$. Igjen setter vi disse uttrykkene inn i den generelle korrelasjonsformelen (1) og får

$$\Gamma = r_s = \frac{\sum\sum (p_j - p_i)(q_j - q_i)}{\sqrt{\sum\sum (p_j - p_i)^2 \sum\sum (q_j - q_i)^2}}.$$

Sammenlikner vi med uttrykket for Pearsons korrelasjon, ser vi at at r_s er ordinær korrelasjon for ranger.

Vi har at $\sum\sum (p_j - p_i)^2 = \sum\sum (q_j - q_i)^2$ siden både p_i og q_i er ranger som går fra 1 til n. Uttrykket for r_s kan forenkles til

$$\Gamma = r_s = \frac{\sum\sum (p_j - p_i)(q_j - q_i)}{\sum\sum (p_j - p_i)^2}.$$

Fordi $\sum p_i$ og $\sum q_j$ er lik summen av de n første positive tallene, kan telleren uttrykkes ved

$$\sum\sum (p_j - p_i)(q_j - q_i) = 2n\sum p_i q_i - 2\sum p_i \sum q_j = 2n\sum p_i q_i - \frac{1}{2}n^2(n+1)^2.$$

Vi definerer $d_i = p_i - q_i$ slik at

$$\sum d_i^2 = \sum (p_i - q_i)^2 = 2\sum p_i^2 - 2\sum p_i q_i.$$

Siden $\sum p_i^2$ er summen av kvadratet av de n første positive tallene, er $\sum p_i^2 = (1/6)n(n+1)(2n+1)$. Det følger at

$$\sum\sum (p_j - p_i)(q_j - q_i) = 2n\sum p_i^2 - \frac{1}{2}n^2(n+1)^2 - n\sum d_i^2 = \frac{1}{6}n^2(n^2-1) - n\sum d_i^2.$$

Nevneren er

$$\sum\sum (p_j - p_i)^2 = 2n\sum p_i^2 - 2\sum\sum p_i p_j = 2n\sum p_i^2 - 2(\sum p_i)^2 = \frac{1}{6}n^2(n^2-1).$$

Innsetting i uttrykket for Γ og noe algebra gir

$$r_s = 1 - \frac{6\sum d_i^2}{n^3 - n}.$$

Parameterversjonen til Spearmans rho betegnes ved ρ_s. Den interesserte leser henvises til Nelsen [38] eller Kruskal [28] for definisjonen av denne.

Felles for alle korrelasjonsmålene er at de kan ta verdier mellom -1 og 1. Når koeffisientenes absoluttverdi nærmer seg 1, øker graden av lineær eller monoton sammenheng mellom to variable. Vi har ingen lineær eller monoton sammenheng dersom koeffisientene har verdien 0. Korrelasjonsmålene er symmetriske om null.

2.2 Valg av test

Pearsons korrelasjonskoeffisient, Kendalls tau og Spearmans rho er beskrivende statistiske mål for graden av sammenheng mellom to eller flere variable. Etter beregning av slike koeffisienter, kan man utføre inferens. Man kan evaluere en eller flere hypoteser angående disse koeffisientene.

Hypoteser som evalueres ved tester basert på Pearsons korrelasjonskoeffisient kan beskrives slik: Finnes det en signifikant *lineær* sammenheng mellom de to variablene i den underliggende populasjonen representert ved utvalget [48]?

For tester basert på Spearmans rho og Kendalls tau vil hypotesen være definert liknende, bortsett fra at vi undersøker om det finnes en signifikant *monoton* sammenheng mellom de to variablene.

Sheskin [48] definerer nullhypotesen til testen basert på Kendalls tau ved

$$H_0 : \tau = 0,$$

hvor τ er definert ved Likning (2) i Seksjon 2.1.2. En tosidig alternativ hypotese uttrykkes ved

$$H_1 : \tau \neq 0.$$

Fordi jeg senere utleder fordelingen til τ når X og Y antas uavhengige, mener jeg en mer presis definisjon av nullhypotesen vil være gitt ved

$$H_0 : X \text{ og } Y \text{ er uavhengige.}$$

Pearsons korrelasjon er passende dersom X og Y har en simultan bivariat normalfordeling. Dersom antakelsen om bivariat normalfordeling ikke er passende, bør man velge en ikke-parametrisk test, eksempelvis Kendalls tau eller Spearmans rho. Ikke-parametriske tester kan være en fordel i flere sammenhenger. Testene er ofte mer robuste fordi de avhenger av færre forutsetninger. De er ofte også mer anvendelige enn parametriske tester. Hvilken test bør foretrekkes av Spearmans rho og Kendalls tau?

Begge testene krever de samme forutsetningene. Forutsetningene vil derfor ikke påvirke valg av test. Vi antar at $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ er uavhengige. I tillegg skal de n parene ha den samme kontinuerlige bivariate fordelingen [38].

For Kendalls tau er a_{ij} og b_{ij} lik ± 1 uansett hvor stor rangavstand det er mellom x_i og x_j eller y_i og y_j. Alle avstander i rang får lik vekt. Dette står i kontrast til Spearmans rho, hvor store rangavstander blir vektet i større grad enn mindre avstander i rang. Valg av metode avhenger av den praktiske situasjonen [27].

Sheskin [48] gir to grunner til at man foretrekker Spearmans rho framfor Kendalls tau. Den ene årsaken er enklere beregninger for r_s. Den andre er at Spearmans rho gir en rimelig god tilnærming til Pearsons korrelasjonskoeffisient når observasjonene er utledet fra en bivariat normalfordeling.

Det finnes også grunner til at en kan foretrekke Kendalls tau. Fordelingen til t konvergerer raskere mot normalfordelingen enn r_s. Normalfordelingen gir en god tilnærming til den eksakte fordelingen til t, selv for små sampelstørrelser [33]. Dessuten er Kendalls tau en forventningsrett estimator for τ. Spearmans rho er ikke en forventningsrett estimator for sin parameterversjon, ρ_s [48].

Til tross for ulikhetene mellom Spearmans rho og Kendalls tau, konkluderer Sheskin [48] og Lindeman et al. [33] at de to observatorene gir mye av den samme informasjonen. De vil, i de fleste tilfeller, resultere i de samme konklusjonene når en tester om den underliggende korrelasjonen er lik null.

2.3 Utledning av den asymptotiske fordelingen til Kendalls tau under nullhypotesen

Med hensyn til den videre oppgaven er det på sin plass å utlede den asymptotiske fordelingen til Kendalls tau under nullhypotesen. Jeg tar utgangspunkt i et eksempel på bruk av sentralgrenseteoremet gitt av Feller [16], samt en bearbeiding av dette eksempelet gitt av Meen

og Heuch [37]. Eksemplene utleder fordelingen til det totale antallet inversjoner i en tilfeldig permutasjon. Jeg setter denne teorien i sammenheng med definisjonen av Kendalls tau. En noe annerledes og springende framstilling av dette beviset er gitt av Walsh [56].

Vi har elementene $(a_1, a_2, \ldots, a_n)$. Disse elementene kan ordnes i $n!$ permutasjoner. Hver av de $n!$ permutasjonene antas like sannsynlige og tildeles derfor sannsynligheten $1/(n!)$. Eksempelvis har $(1, 2, 3)$ 6 permutasjoner: $(1, 2, 3), (1, 3, 2), (2, 1, 3), (2, 3, 1), (3, 1, 2)$ og $(3, 2, 1)$. I en gitt permutasjon sier man at elementet a_k produserer r inversjoner dersom det står nøyaktig r elementer med indeks mindre enn k foran a_k. Denne definisjonen er gitt av Meen og Heuch [37]. Dersom vi har permutasjonen (a_2, a_1, a_4, a_3), produserer elementene a_1 og a_2 ingen inversjoner, mens både a_3 og a_4 produserer to hver. Totalt blir det fire inversjoner.

Feller [16] definerer inversjoner annerledes. I en gitt permutasjon produserer elementet a_k r inversjoner dersom dette elementet står foran nøyaktig r elementer med lavere indeks. Med en slik definisjon vil man med permutasjonen (a_2, a_1, a_4, a_3) si at elementet a_1 produserer null inversjoner, a_2 produserer en, a_3 ingen, mens a_4 produserer en. I alt får en to inversjoner. Dette kan ved første øyekast virke merkelig. Den første definisjonen ser på inversjoner sett i forhold til $(a_n, a_{n-1}, \ldots, a_2, a_1)$. Den sistnevnte ser på inversjoner sett i forhold til $(a_1, a_2, \ldots, a_{n-1}, a_n)$. Hvis ikke annet er nevnt, nytter jeg definisjonen til Meen og Heuch.

Det totale antallet inversjoner i en tilfeldig permutasjon betegnes S_n og defineres ved $S_n = X_1 + X_2 + \ldots + X_n$. Her er X_k antall inversjoner produsert av a_k for $1 \leq k \leq n$, sett i forhold til $(a_n, \ldots, a_2, a_1)$. Det presiseres at X ikke er den samme variabelen som tidligere i Kapittel 2.

Den stokastiske variabelen X_k kan ta verdiene $0, 1, \ldots, k-1$. Hver enkelt verdi har sannsynlighet lik $1/k$ [16]. Vi har altså en diskret uniform fordeling over $0, 1, \ldots, k-1$.

Jeg finner forventningen og variansen til X_k. Enkel regning gir

$$\begin{aligned} \mathrm{E}(X_k) &= \frac{1}{k}\sum_{i=0}^{k-1} i = \frac{1}{k}\sum_{i=0}^{k} i - \frac{1}{k}k \\ &= \frac{1}{k}\frac{k(k+1)}{2} - 1 = \frac{k+1}{2} - 1 \\ &= \frac{k+1-2}{2} = \frac{k-1}{2}. \end{aligned}$$

Videre er

$$\begin{aligned}\mathrm{Var}(X_k) &= \mathrm{E}(X_k^2) - \mathrm{E}(X_k)^2 = \frac{1}{k}\sum_{i=0}^{k-1} i^2 - \left(\frac{k-1}{2}\right)^2 \\ &= \frac{1}{k}\sum_{i=0}^{k} i^2 - \frac{1}{k}k^2 - \frac{k^2-2k+1}{4} = \frac{k(k+1)(2k+1)}{6k} - k - \frac{k^2-2k+1}{4} \\ &= \frac{2k^2+3k+1}{6} - k - \frac{k^2-2k+1}{4} = \frac{4k^2+6k+2-12k-3k^2+6k-3}{12} \\ &= \frac{k^2-1}{12}.\end{aligned}$$

Forventningen til S_n finnes ved å nytte uttrykket til forventningen til X_k. Det følger at

$$\begin{aligned}\mathrm{E}(S_n) &= \mathrm{E}(X_1 + X_2 + \cdots + X_n) \\ &= \sum_{k=1}^{n} \mathrm{E}(X_k) = \sum_{k=1}^{n} \frac{k-1}{2} = \frac{1}{2}\sum_{k=1}^{n} k - \frac{1}{2}\sum_{k=1}^{n} 1 \\ &= \frac{1}{2}\frac{n(n+1)}{2} - \frac{1}{2}n = \frac{n^2}{4} + \frac{n}{4} - \frac{2n}{4} \\ &= \frac{n(n-1)}{4}.\end{aligned}$$

Antall inversjoner produsert av a_k er uavhengig av hvordan $a_1, a_2, \ldots, a_{k-1}$ er ordnet innbyrdes. Det følger at $X_1, X_2, \ldots, X_k$ er uavhengige stokastiske variable [16]. Variansen til S_n er gitt ved

$$\begin{aligned}\mathrm{Var}(S_n) &= \mathrm{Var}(X_1 + X_2 + \cdots + X_n) \\ &= \sum_{k=1}^{n} \mathrm{Var}(X_k) = \sum_{k=1}^{n} \frac{k^2-1}{12} = \frac{1}{12}\sum_{k=1}^{n} k^2 - \frac{1}{12}\sum_{k=1}^{n} 1 \\ &= \frac{1}{12}\frac{n(n+1)(2n+1)}{6} - \frac{1}{12}n = \frac{1}{12\cdot 6}(n^2+n)(2n+1) - \frac{1}{12}\cdot n \\ &= \frac{2n^3+n^2+2n^2+n}{72} - \frac{1}{12}\cdot n \\ &= \frac{2n^3+3n^2-5n}{72}.\end{aligned}$$

Hvordan kan en nytte informasjon om S_n til å utlede fordelingen til Kendalls tau under nullhypotesen? Kendalls tau er definert ved

$$t = \frac{2(C-D)}{n(n-1)} = \frac{2S}{n(n-1)},$$

hvor $S = C - D$. Som tidligere betegner C antall konkordante par, mens D står for antall diskordante par. Jeg tar utgangspunkt i et eksempel gitt av Kendall og Gibbons [27] og viser hvordan en effektivt kan finne antall konkordante par.

Ti gutter er rangert etter deres evne i matematikk og musikk.

Gutt:	A	B	C	D	E	F	G	H	I	J
Matematikk:	7	4	3	10	6	2	9	8	1	5
Musikk:	5	7	3	10	1	9	6	2	8	4

Finnes det en sammenheng mellom evner i matematikk og musikk? Forfatterne ordner matematikkegenskapene slik at de står i den naturlige rekkefølgen.

Gutt:	I	F	C	B	J	E	A	H	G	D
Matematikk:	1	2	3	4	5	6	7	8	9	10
Musikk:	8	9	3	7	4	1	5	2	6	10

For å finne antall konkordante par er det nok kun å se på den nåværende rangeringen av musikkegenskapene. Vi ser først på tallet 8. Ingen tall med lavere indeks står til venstre for dette tallet. Antall konkordante par er foreløpig null. Fordi 8 står til venstre for 9, vil vi få et bidrag til C på $+1$. Vi fortsetter på denne måten og finner at $C = 0+1+0+1+1+0+3+1+5+9 = 21$. Dersom vi nytter Meen og Heuchs definisjon av inversjoner, tilsvarer antall konkordante par, C, det totale antallet inversjoner i en tilfeldig permutasjon, S_n. Igjen vil jeg presisere at vi er under nullhypotesen. Jeg antar at alle permutasjoner av musikkegenskapene er like sannsynlige. Jeg forestiller meg at eksempelet ovenfor viser en tilfeldig permutasjon.

Vi har at $C + D = 2^{-1}n(n-1) = \binom{n}{2}$, antall måter en kan velge to objekter ut fra n objekter dersom en ikke tillater tilbakelegging og trekningens rekkefølge er uten betydning. Det følger at $D = \binom{n}{2} - C = \binom{n}{2} - S_n$.

Ved bruk av Meen og Heuchs definisjon av inversjoner får en at

$$S = C - D = S_n - \left(\binom{n}{2} - S_n\right) = 2S_n - \binom{n}{2}.$$

Forventningen og variansen til S er gitt ved

$$\begin{aligned}\mathrm{E}(S) &= \mathrm{E}\left(2S_n - \binom{n}{2}\right) = 2\mathrm{E}(S_n) - \frac{1}{2}n(n-1)\\ &= \frac{2n(n-1)}{4} - \frac{1}{2}n(n-1) = 0.\end{aligned}$$

og

$$\begin{aligned}\mathrm{Var}(S) &= \mathrm{Var}\left(2S_n - \binom{n}{2}\right) = \mathrm{Var}(2S_n) = 4\mathrm{Var}(S_n)\\ &= 4\cdot\frac{2n^3+3n^2-5n}{72} = \frac{2n^3+3n^2-5n}{18}\\ &= \frac{n(2n^2+3n-5)}{18} = \frac{n(n-1)(2n+5)}{18}.\end{aligned}$$

Jeg oppnår de samme resultatene om jeg nytter Fellers definisjon av inversjoner. La $S_n = X_1 + X_2 + \ldots + X_n$ betegne det totale antallet inversjoner i en tilfeldig permutasjon, sett i forhold til permutasjonen $(a_1, a_2, \ldots, a_n)$. Med denne definisjonen vil S_n tilsvare D og ikke C. Walsh [56] nytter denne framgangsmåten.

Forventningen og variansen til t følger lett, og vi får at

$$\mathrm{E}(t) = \mathrm{E}\left(\frac{2S}{n(n-1)}\right) = \frac{2}{n(n-1)}\mathrm{E}(S) = 0$$

og

$$\mathrm{Var}(t) = \mathrm{Var}\left(\frac{2S}{n(n-1)}\right) = \frac{4}{n^2(n-1)^2}\mathrm{Var}(S) = \frac{2(2n+5)}{9n(n-1)}.$$

Det gjenstår å vise at t er asymptotisk normalfordelt. Til dette bruker jeg Lindebergs sentralgrenseteorem som definert hos Meen og Heuch [37]. Vi vet at $X_1, X_2, \ldots, X_n$ er uavhengige stokastiske variable med endelig forventning og varians. Videre er $F_k(x)$ fordelingsfunksjonen til X_k. Som tidligere defineres $S_n = X_1 + X_2 + \cdots + X_n$, hvor det totale antallet inversjoner er sett i forhold til permutasjonen $(a_n, \ldots, a_2, a_1)$. Vi betrakter de standardiserte summene

$$Z_n = \left(S_n - \sum_{k=1}^{n} \mathrm{E}(X_k)\right) / \sqrt{\mathrm{Var}(S_n)}, \quad n = 1, 2, \ldots \quad .$$

Lindebergbetingelsen er oppfylt dersom

$$\lim_{n\to\infty} \Big\{ \frac{1}{\mathrm{Var}(S_n)} \sum_{k=1}^{n} \int\limits_{|x-\mathrm{E}(X_k)|>\delta\sqrt{\mathrm{Var}(S_n)}} \big(x - \mathrm{E}(X_k)\big)^2 dF_k(x) \Big\} = 0 \quad \text{for alle } \delta > 0.$$

Hvis Lindebergbetingelsen er oppfylt, gjelder det at

$$Z_n \xrightarrow{L} Z \sim N(0,1),$$

hvor $\xrightarrow{L}$ betegner konvergens i fordeling. Meen og Heuch [37] gir et bevis for Lindebergs sentralgrenseteorem.

I vår situasjon er

$$\mathrm{Var}(S_n) = \frac{2n^3 + 3n^2 - 5n}{72} \sim \frac{n^3}{36}.$$

For alle $\delta > 0$ er $\delta\sqrt{\mathrm{Var}(S_n)} > n > n/2$ så lenge n er stor nok. Vi får at

$$\int\limits_{|x-\mathrm{E}(X_k)|>\delta\sqrt{\mathrm{Var}(S_n)}} \big(x - \mathrm{E}(X_k)\big)^2 dF_k(x) = 0, \quad k = 1, 2, \ldots, n.$$

Denne likheten gjelder fordi $F_k(x)$ er konstant for $x < 0$ og $x \geq k-1$. Når $0 \leq x < k-1$, vil $|x - \mathrm{E}(X_k)| < n/2 < \delta\sqrt{\mathrm{Var}(S_n)}$. Lindebergbetingelsen er oppfylt, og vi har at

$$Z_n = \frac{S_n - \mathrm{E}(S_n)}{\sqrt{\mathrm{Var}(S_n)}} \xrightarrow{L} N(0,1).$$

Jeg mener Meen og Heuch [37] er upresise i deres begrunnelse for at det overnevnte integralet er lik 0. De sier $F_k(x)$ er konstant for $x < 0$ og $x > k$. Forfatterne gjør ikke rede for hva som hender når $x = k$. Hvis $x = k = n$, følger det at $\mathrm{E}(X_n) = (n-1)/2$. Videre er $|x - \mathrm{E}(X_k)| = |n - (n-1)/2| = |(n+1)/2| > n/2$. Det er derfor et poeng ved å presisere at $F_k(x)$ er konstant også når $x = k$.

Fordi S er et lineært uttrykk av S_n, er observatoren asymptotisk normalfordelt med forventning lik 0 og varians gitt ved $n(n-1)(2n+5)/18$. Den asymptotiske normalfordelingen følger fra Helly-Slutskys setning, definert hos Meen og Heuch [37]. Også t er et lineært uttrykk av S_n. Følgelig er t asymptotisk normalfordelt. Jeg har bevist at

$$t \sim N\left(0\ ,\ \frac{2(2n+5)}{9n(n-1)}\right).$$

Under den alternative hypotesen vil fordelingen til Kendalls tau forandre seg noe. Vi vil fortsatt ha normalfordeling, men dette følger ikke direkte fra beviset i denne seksjonen. Forventningen og variansen vil ikke forbli den samme.

Dersom forutsetningen for å utføre en rangkorrelasjonstest brytes, eller dersom det finnes en monoton sammenheng mellom X og Y under nullhypotesen, vil jeg ikke kunne nytte dette beviset til å si noe om den asymptotiske fordelingen til tau. Innledningsvis antar jeg at de $n!$ permutasjonene av elementene $(a_1, a_2, \ldots, a_n)$ er like sannsynlige. Allerede her bryter beviset sammen.

2.4 Enda en utledning av variansen til Kendalls tau under nullhypotesen

Jeg har bevist det jeg skulle, men ønsker likevel å inkludere enda en utledning av variansen til Kendalls tau. Jeg tar utgangspunkt i beviset gitt av Kendall og Gibbons [27]. Beviset er vanskelig å forstå slik det er forklart i denne boken. Jeg forsøker å forenkle det.

Først er det nødvendig med noen definisjoner. Definisjonene inngår også tidligere i kapittelet om Kendalls tau. Jeg gjentar dem med noe ulik notasjon slik at det skal være lettere å følge beviset.

Vi har observasjonsparene $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Som før defineres

$$t = \frac{2S}{n(n-1)}.$$

Observatoren S kan beregnes ved $S = \sum_{i<j} a_{ij} b_{ij}$, hvor

$$a_{ij} = \mathrm{sgn}(x_j - x_i) = \begin{cases} 1 & \text{hvis } x_i < x_j \\ 0 & \text{hvis } x_i = x_j \\ -1 & \text{hvis } x_i > x_j \end{cases}$$

og

$$b_{ij} = \text{sgn}(y_j - y_i) = \begin{cases} 1 & \text{hvis } y_i < y_j \\ 0 & \text{hvis } y_i = y_j \\ -1 & \text{hvis } y_i > y_j \end{cases}.$$

Kendall og Gibbons [27] innfører $c_{ij} = a_{ij}b_{ij}$ og $c = \sum_{i,j}^{n} c_{ij}$, slik at $c = 2S$. Jeg ønsker å finne $\text{Var}(t)$ og starter med å finne $\text{Var}(c)$. Vi har at $\text{Var}(c) = \text{E}(c^2) - \text{E}(c)^2$. Fordi a_{ij} uttrykkes kun ved hjelp av X_i og X_j, b_{ij} uttrykkes kun ved hjelp av Y_i og Y_j, og X og Y er uavhengige under nullhypotesen, har vi at

$$\text{E}(c) = \text{E}\sum_{i,j}^{n}(c_{ij}) = \text{E}\sum_{i,j}^{n}(a_{ij}b_{ij}) = \sum_{i,j}^{n}\text{E}(a_{ij}b_{ij}) = \sum_{i,j}^{n}\text{E}(a_{ij})\text{E}(b_{ij}).$$

For å holde meg nær notasjonen til Kendall og Gibbons [27], tillater jeg noe misbruk av notasjon. Det er underforstått at a_{ij}, b_{ij}, c_{ij} og c er stokastiske variable når jeg utfører inferens. Symmetri gir at

$$\text{E}(a_{ij}) = 1\cdot P(a_{ij} = 1) + 0\cdot P(a_{ij} = 0) + (-1)\cdot P(a_{ij} = 1) = \frac{1}{2} - \frac{1}{2} = 0.$$

Sannsynligheten for at $x_i = x_j$ er lik null fordi X er kontinuerlig fordelt. Følgelig er både $\text{E}(c) = 0$ og $\text{E}(S) = 0$.

Neste steg er å finne $\text{E}(c^2)$. Vi har

$$\text{E}(c^2) = \text{E}\big(\sum_{i,j}^{n} a_{ij}b_{ij}\big)^2 = \text{E}\big(\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}b_{ij}\sum_{k=1}^{n}\sum_{l=1}^{n} a_{kl}b_{kl}\big).$$

Hvordan går jeg videre herfra? I likhet med Kendall og Gibbons [27] sorterer jeg de ulike leddene i summasjonsuttrykket. Jeg teller opp og samler ledd med lik forventning. Deretter gjenstår det kun å finne forventningen til de ulike leddene.

Jeg gjennomgår sorteringen først. I utgangspunktet har vi n^4 ledd. Jeg ekskluderer ledd som ikke gir bidrag, det vil si ledd som har verdien null. Da gjenstår $(n^2 - n)(n^2 - n) = n^2(n-1)^2$ ledd. Disse leddene kan deles opp i tre hovedtyper. Vi har ledd der alle indeksene er ulike. Dette betegner jeg $i \neq j \neq k \neq l$, selv om denne notasjonen muligens kan være villedende. Vi har også ledd av typen $a_{ij}a_{ik}b_{ij}b_{ik}$, altså ledd der to av indeksene er like. Sist, men ikke minst, har vi ledd av typen $a_{ij}^2 b_{ij}^2$. Her er to og to indekser like. Igjen må en legge merke til at vi ekskluderer ledd som ikke gir bidrag, eksempelvis ledd av typen $a_{ii}b_{ii}a_{kl}b_{kl}$. Dette er den samme inndelingen som brukes av Kendall og Gibbons [27].

Vi har $n(n-1)(n-2)(n-3)$ ledd der alle indeksene er ulike.

Det er $n(n-1)(n-2)$ måter å trekke tre forskjellige indekser på ut av n mulige. Når vi ekskluderer ledd som ikke gir bidrag, vil vi ha ledd av typen $a_{ij}a_{ik}b_{ij}b_{ik}$, $a_{ji}a_{ik}b_{ji}b_{ik}$, $a_{ij}a_{ki}b_{ij}b_{ki}$ og $a_{ji}a_{ki}b_{ji}b_{ki}$ [27].

Til slutt er det $n(n-1)$ mulige måter å trekke to forskjellige indekser på ut av n. Vi har leddkombinasjonene $a_{ij}^2 b_{ij}^2$ og $a_{ij}b_{ij}a_{ji}b_{ji}$.

Dersom forutsetningene er oppfylt under nullhypotesen, kan vi anta at $P(a_{ij}b_{ij} = q) = P(a_{kl}b_{kl} = q)$ for $q = \pm 1$, $i \neq j$ og $k \neq l$. Dette gir at $\mathrm{E}(a_{ij}b_{ij}) = \mathrm{E}(a_{kl}b_{kl})$. Benytter vi dette, får vi at

$$\begin{aligned}\mathrm{Var}(c) = \mathrm{E}(c^2) &= n(n-1)(n-2)(n-3)E(a_{ij}b_{ij}a_{kl}b_{kl}) \\ &\quad + 4n(n-1)(n-2)\mathrm{E}(a_{ij}a_{ik}b_{ij}b_{ik}) + n(n-1)\mathrm{E}(a_{ij}^2 b_{ij}^2) + n(n-1)\mathrm{E}(a_{ij}b_{ij}a_{ji}b_{ji}).\end{aligned}$$

Opptellingen er foretatt. Det gjenstår å finne de fire forventningene.

Under nullhypotesen vet vi at $\mathrm{E}(a_{ij}b_{ij}a_{kl}b_{kl}) = \mathrm{E}(a_{ij}a_{kl})\mathrm{E}(b_{ij}b_{kl})$. Dersom testens forutsetninger er oppfylt og alle indeksene er ulike under nullhypotesen, vet vi at X_i, X_j, X_k, X_l er uavhengige. Da er $\mathrm{E}(a_{ij}a_{kl}) = \mathrm{E}(a_{ij})\mathrm{E}(a_{kl}) = 0$. Første ledd i uttrykket ovenfor forsvinner. Her er min metode ulik metoden til Kendall og Gibbons [27]. For å demonstrere at $\mathrm{E}\big(\sum_{i\neq j\neq k\neq l} a_{ij}b_{ij}a_{kl}b_{kl}\big)$ forsvinner, mener forfatterne det nok å vise at $\mathrm{E}\big(\sum_{i\neq j\neq k\neq l} a_{ij}a_{kl}\big) = 0$. Dette synes ikke jeg er intuitivt. Vi kan ikke uten videre anta at

$$\mathrm{E}\Big(\sum_{i\neq j\neq k\neq l} a_{ij}b_{ij}a_{kl}b_{kl}\Big) = \mathrm{E}\Big(\sum_{i\neq j\neq k\neq l} a_{ij}a_{kl}\sum_{i\neq j\neq k\neq l} b_{ij}b_{kl}\Big) = \mathrm{E}\Big(\sum_{i\neq j\neq k\neq l} a_{ij}a_{kl}\Big)\mathrm{E}\Big(\sum_{i\neq j\neq k\neq l} b_{ij}b_{kl}\Big).$$

Under nullhypotesen kan forventningen til $a_{ij}a_{ik}b_{ij}b_{ik}$ skrives som $\mathrm{E}(a_{ij}a_{ik})\mathrm{E}(b_{ij}b_{ik})$. Forventningen til $a_{ij}a_{ik}$ er lik forventningen til $b_{ij}b_{ik}$ dersom forutsetningene er oppfylt. Vi har at $\mathrm{E}(a_{ij}a_{ik}) = P(a_{ij}a_{ik} = 1) - P(a_{ij}a_{ik} = -1) = \frac{2}{3} - \frac{1}{3} = \frac{1}{3} = \mathrm{E}(b_{ij}b_{ik})$. Dette krever et bevis. Beviset er nokså langt. Av den grunn følger det i Seksjon 2.5. Jeg savner et bevis av denne typen i utledningen til Kendall og Gibbons [27].

Fordi a_{ij} og b_{ij} kun kan ta verdien -1 eller 1 dersom $i \neq j$, har vi at $\mathrm{E}(a_{ij}^2 b_{ij}^2) = \mathrm{E}(a_{ij}^2)\mathrm{E}(b_{ij}^2) = 1$. Videre er $\mathrm{E}(a_{ij}b_{ij}a_{ji}b_{ji}) = \mathrm{E}(a_{ij}a_{ji})\mathrm{E}(b_{ij}b_{ji}) = (-1)(-1) = 1$. Kendall og Gibbons [27] behandler disse to uttrykkene under ett. Jeg velger å behandle dem hver for seg, da framgangsmåten for å finne forventningen er noe ulik.

Innsetting av forventningsuttrykkene i $\mathrm{Var}(c)$ gir

$$\begin{aligned}\mathrm{Var}(c) &= n(n-1)(n-2)(n-3)\cdot 0 \\ &\quad + 4n(n-1)(n-2)\cdot\frac{1}{3}\cdot\frac{1}{3} + 2n(n-1)\cdot 1\cdot 1 \\ &= 2n(n-1)\left(\frac{2}{9}(n-2)+1\right) = 2n(n-1)\left(\frac{2n-4+9}{9}\right) \\ &= \frac{2}{9}n(n-1)(2n+5).\end{aligned}$$

Videre er

$$\mathrm{Var}(S) = \mathrm{Var}\left(\frac{c}{2}\right) = \frac{1}{4}\mathrm{Var}(c) = \frac{n(n-1)(2n+5)}{18}.$$

Uttrykket er det samme som jeg kom fram til i mitt forrige bevis. Igjen har jeg bevist at

$$\mathrm{Var}(t) = \frac{2(2n+5)}{9n(n-1)}.$$

Jeg unngår vanskelige summasjonsuttrykk i utledningen av variansen til Kendalls tau. Dette forenkler beregningene betraktelig sammenliknet med beviset til Kendall og Gibbons.

Dersom vi ikke er under nullhypotesen, vil variansen bli vanskeligere å utlede. Flere steder nytter jeg at $\mathrm{E}(ab) = \mathrm{E}(a)\mathrm{E}(b)$. Dette kan kun gjøres dersom a og b er uavhengige. Eksempelvis må forventningen til c modifiseres dersom vi antar at $\mathrm{Cov}(a_{ij}, b_{ij}) = \rho$. Vi får at

$$\mathrm{E}(c) = \sum_{i,j}^{n} \mathrm{E}(a_{ij}b_{ij}) = \sum_{i,j}^{n} \mathrm{Cov}(a_{ij}, b_{ij}) = n(n-1)\rho.$$

Ledd som ikke gir bidrag er ekskludert. Dette gir $\mathrm{E}(t) = \rho$.

Jeg vil kort påpeke enkelte faremomenter ved beviset dersom forutsetningene for å utføre rangkorrelasjonstesten basert på Kendalls tau ikke er oppfylt. Dersom observasjonsparene er avhengige, må uttrykkene for $\mathrm{E}(a_{ij}b_{ij}a_{kl}b_{kl})$ og $\mathrm{E}(a_{ij}b_{ij}a_{ik}b_{ik})$ modifiseres. Vi kan eksempelvis ikke skrive $\mathrm{E}(a_{ij}a_{kl}) = \mathrm{E}(a_{ij})\mathrm{E}(a_{kl})$ dersom X-variablene ikke er uavhengige.

Jeg har også tidligere antatt at $P(a_{ij} = 1) = P(a_{ij} = -1) = 1/2$. Dette kan ikke lenger antas dersom de observerte verdiene av X ikke stammer fra den samme kontinuerlige fordelingen. I dette tilfellet kan vi heller ikke anta at $\mathrm{E}(a_{ij}) = \mathrm{E}(a_{kl})$.

Noether [40] gir et tredje bevis for $\mathrm{Var}(S)$. Han tar utgangspunkt i

$$\mathrm{Var}(S) = \mathrm{Var}\big(\sum_{i<j} a_{ij}b_{ij}\big) = \sum_{i<j} \mathrm{Var}(a_{ij}b_{ij}) + \sum_{i<j}\sum_{k<l} \mathrm{Cov}(a_{ij}b_{ij}, a_{kl}b_{kl}),$$

og arbeider videre med disse uttrykkene.

2.5 Forventningen til $a_{ij}a_{ik}$

I Kendall og Gibbons bevis for den asymptotiske variansen til tau må en vite at $E(a_{ij}a_{ik}) = 1/3$. Jeg utleder denne forventningen.

Vi har $n(n-1)(n-2)$ ledd av typen $a_{ij}a_{ik}$, hvor $i \neq j \neq k$. Alle permutasjoner er like sannsynlige. Hvor mange av disse leddene oppfyller $a_{ij}a_{ik} = -1$? Fra definisjonen av a_{ij} ser vi at $a_{ij}a_{ik} = -1$ kun dersom $X_j < X_i$ samtidig som $X_k > X_i$, eller hvis $X_j > X_i$ samtidig som $X_k < X_i$. Disse hendelsene er disjunkte. Jeg tar for meg det første tilfellet først og teller hvor mange av de $n(n-1)(n-2)$ leddene som tilfredsstiller $X_j < X_i$ og $X_k > X_i$.

La $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ betegne de ordnede verdiene av $X_1, X_2, \ldots, X_n$. Når $X_i = X_{(1)}$, vil aldri $X_j < X_i$ og samtidig $X_k > X_i$. Når $X_i = X_{(2)}$, må $X_j = X_{(1)}$ samtidig som X_k må ha ordning større enn 2. Det finnes $n-2$ ledd hvor dette er oppfylt. Dersom $X_i = X_{(3)}$,

må j være ordnet som nummer 1 eller 2, mens X_k må ha en ordning større enn 3. Det er $2(n-3)$ ledd av denne typen. Jeg kaller disse for gunstige ledd.

Slik kan man fortsette å telle, og en kan sette opp følgende skjema:

Ordning til X_i	Antall gunstige ledd
1	0
2	$1\cdot(n-2)$
3	$2\cdot(n-3)$
4	$3\cdot(n-4)$
$\vdots$	$\vdots$
$n-1$	$(n-2)\cdot 1$
n	$(n-1)\cdot 0$

Jeg summerer og kommer fram til en formel for antall ledd hvor $X_j < X_i$ og $X_k > X_i$. Vi får at

$$\begin{aligned}\sum_{l=0}^{n-1} l\cdot(n-(l+1)) &= \sum_{l=0}^{n-1} l\cdot(n-1) - \sum_{l=0}^{n-1} l^2 = (n-1)\sum_{l=0}^{n-1} l - \sum_{l=0}^{n-1} l^2 \\ &= \frac{(n-1)^2 n}{2} - \frac{(n-1)n\big(2(n-1)+1\big)}{6} \\ &= \frac{n(n-1)}{6}\big(3(n-1)-(2(n-1)+1)\big) \\ &= \frac{n(n-1)(n-2)}{6}.\end{aligned}$$

Tilsvarende kan en telle hvor mange ledd som oppfyller $X_i < X_j$ og $X_i > X_k$. Vi får den samme formelen. Antall ledd som oppfyller $a_{ij}a_{ik} = -1$ er derfor lik $(2/6)n(n-1)(n-2)$.

Vi finner at

$$P(a_{ij}a_{ik} = -1) = \frac{\text{antall gunstige}}{\text{antall mulige}} = \frac{2}{6}\frac{n(n-1)(n-2)}{n(n-1)(n-2)} = \frac{1}{3}.$$

Verdiene ± 1 utgjør hele utfallsrommet til $a_{ij}a_{ik}$. Jeg vet derfor at

$$P(a_{ij}a_{ik} = 1) = 1 - \frac{1}{3} = \frac{2}{3}.$$

Forventningen er gitt ved

$$E(a_{ij}a_{ik}) = P(a_{ij}a_{ik} = 1) - P(a_{ij}a_{ik} = -1) = \frac{2}{3} - \frac{1}{3} = \frac{1}{3},$$

som er hva jeg ønsket å vise.

3 Beskrivelse av Begg og Mazumdars testmetode og simuleringssituasjoner

3.1 Motivasjon for testmetoden

Dersom en metaanalyse ikke inneholder publikasjonsbias, vil funnelplottet, med studiers sampelstørrelse langs den vertikale aksen og effektestimat langs den horisontale, likne en omvendt symmetrisk trakt [15]. En forventer at studienes sampelstørrelse er ukorrelert med effektestimatene. Hvis en metaanalyse inneholder publikasjonsbias, vil funnelplottet i de fleste tilfeller være asymmetrisk. Studienes sampelstørrelse og effektestimat er da korrelerte. Hvis effektestimatene måles i log-odds-ratio og odds-ratio antas å være større enn eller lik en, vil denne korrelasjonen være negativ. Korrelasjonen er positiv om odds-ratio antas å være mindre enn eller lik en.

Negativ korrelasjon mellom studiers sampelstørrelse og effektestimat impliserer positiv korrelasjon mellom effektestimatene og deres varianser [8]. Positiv korrelasjon mellom studiers sampelstørrelse og effektestimat impliserer negativ korrelasjon mellom effektestimatene og deres varianser.

Graden av korrelasjon kan undersøkes ved hjelp av statistiske tester. Med dette som utgangspunkt introduserer Begg og Mazumdar [8] en rangkorrelasjontest basert på Kendalls tau for å identifisere eventuell publikasjonsbias i metaanalyser.

3.2 Testmetoden til Begg og Mazumdar

En metaanalyse inneholder k studier. De estimerte effektstørrelsene betegnes $t_1, t_2, \ldots, t_k$. Sampelvariansene betegnes $v_1, v_2, \ldots, v_k$. Rangkorrelasjonstesten er basert på antakelsen om at effektstørrelsene er uavhengige og identisk fordelte under nullhypotesen om ingen publikasjonsbias. Effektestimatene bør derfor standardiseres før en kan utføre testen [5].

Begg og Mazumdar [8] korrelerer t_i^* og v_i for $i = 1, 2, \ldots, k$ ved hjelp av Kendalls tau, hvor

$$t_i^* = \frac{t_i - \bar{t}}{(v_i^*)^{1/2}}.$$

Forfatterne definerer

$$\bar{t} = \frac{\sum v_j^{-1} t_j}{\sum v_j^{-1}}$$

og

$$v_i^* = v_i - \big(\sum v_j^{-1}\big)^{-1},$$

hvor v_i^* er variansen til $t_i - \bar{t}$.

Vi tester nullhypotesen om at t_i^* og v_i er uavhengige for $i = 1, 2, \ldots, k$. Hvis en forkaster nullhypotesen om uavhengighet, forkaster en også nullhypotesen om ingen publikasjonsbias. Den normaliserte testobservatoren defineres ved

$$z = \frac{C - D}{\sqrt{k(k-1)(2k+5)/18}}.$$

Dette uttrykket er i henhold til resultatene om Kendalls tau fra det forrige kapittelet. Nevneren bør modifiseres dersom vi har ties. Jeg vil ikke utdype dette nærmere.

3.3 Faktorer som kan påvirke testens styrke

Flere faktorer kan påvirke testens styrke. Begg og Mazumdar [8] undersøker styrken i forhold til faktorene som nevnes i denne seksjonen ved hjelp av simuleringer.

Sampelstørrelse er en faktor som er kjent for å påvirke testers styrke. Det undersøkes hvordan antall studier i metaanalysen, k, vil påvirke styrken. Legg merke til skillet mellom en metaanalyses sampelstørrelse og en studies sampelstørrelse. Sampelstørrelsen til en metaanalyse er antall studier i denne analysen. En studies sampelstørrelse refererer til antall objekter som deltar i studien. Denne studien inngår i en metaanalyse som én av totalt k studier.

Vi har en modell med faste effekter. Det undersøkes hvordan den underliggende effektparameteren, δ, påvirker styrken. Når δ beveger seg bort fra nullverdien, vil seleksjonspresset minke. Færre og færre studier ekskluderes fra metaanalysene. Vi forventer derfor at styrken vil minke med økende verdier av δ.

Testen er basert på det faktum at seleksjonseffekten er ulik for studier med liten sampelstørrelse og stor varians og studier med stor sampelstørrelse og liten varians. Hvis alle studiene har omtrent den samme variansen, vil det være vanskelig å oppdage publikasjonsbias. Det forventes at stor variasjon i variansene vil gi økt styrke sammenliknet med liten variasjon.

Forfatterne undersøker videre hvordan seleksjonsstyrken påvirker testens styrke. De bruker forskjellige seleksjonsmodeller, se Seksjon 3.4.1. Er testen robust i forhold til valg av seleksjonsmodell? Begg og Mazumdar vurderer både ensidig og tosidig seleksjon. Med ensidig seleksjon mener man at studier med positiv effekt, effekt i ønsket retning, oftere blir publisert enn studier med negativ effekt.

3.4 Simuleringer

Et første naturlig steg i prosessen for å forstå og analysere deler av Begg og Mazumdars artikkel [8] er å kontrollere artikkelens simuleringsresultater.

I denne seksjonen beskrives simuleringsprosedyren som Begg og Mazumdar bruker for å undersøke testmetodens egenskaper. Det skal være tilstrekkelig å lese denne seksjonen for å

kunne gjennomføre tilsvarende simuleringer. Dersom forfatterne kommenterer trekk i framgangsmåten som er sentrale for å forstå videre diskusjoner i min oppgave, vil disse kort bli nevnt. Jeg vil derimot ikke kommentere eller utdype trekk ved simuleringen som ikke har betydning for videre lesing. Jeg ser det ikke som min oppgave å forsvare de valgene forfatterne foretar, men heller å arbeide videre ut fra de valgene som allerede er tatt. Ønskes dypere innsikt og forståelse for forfatternes parametervalg og valg av seleksjonsmodeller, henvises leseren til Begg og Mazumdars artikkel [8].

3.4.1 Seleksjonsmodeller

Effektestimatet, t_i, genereres tilfeldig fra en normalfordeling med forventning δ og varians v_i. Studiene er designet slik at de estimerer en felles effektstørrelse, δ. Variansene avhenger av sampelstørrelsen i de individuelle studiene [8].

Effektestimatet bygger på bidrag fra mange observasjoner. Grunnet sentralgrenseteoremet vil effektestimatet, etter en passende transformasjon, i de fleste tilfeller ha en asymptotisk normalfordeling. Antakelsen om normalfordelte effektestimat har derfor den fordelen at resultatene gjelder på et mer generelt grunnlag. Dersom studienes sampelstørrelse er stor nok, vil resultatene gjelde uavhengig om effektestimatet måles i eksempelvis odds-ratio, relativ risiko eller hazard-ratio. Vedlegg B omhandler odds-ratio.

Publikasjonsbias simuleres ved hjelp av vektfunksjoner. En vektfunksjon gir sannsynligheten for om en studie blir publisert. Begg og Mazumdar [8] bruker ulike vektfunksjoner. Den første avhenger av den observerte p-verdien for hypotesen om at studiens effekt er lik null. Vektfunksjonen defineres ved

$$w_i\big(t_i(p_i)\big) = s(p_i) = \exp(-bp_i^a),$$

hvor $s(p_i)$ er vektfunksjonen beregnet i $p_i = \Phi(-t_i/v_i^{1/2})$. Seleksjonen er ensidig. Vektfunksjonen er en monotont synkende funksjon av den ensidige p-verdien. Den andre vektfunksjonen avhenger av det observerte effektestimatet, t_i, og defineres ved

$$w(t_i) = \exp\big(-b\Phi(-t_i)^a\big).$$

Begg og Mazumdar vurderer også en tredje mulighet. De nytter en vektfunksjon som er lik den førstnevnte, med unntak av at $p_i = 2\Phi(-|t_i|/v_i^{1/2})$. Legg merke til trykkfeil i artikkelen til Begg og Mazumdar, hvor p_i defineres som $p_i = 2\Phi(-t_i/v^{1/2})$. Dette er en tosidig test. Begg [5] gir en logisk begrunnelse for å vurdere seleksjonsmodeller basert på p-verdi og seleksjonsmodeller basert på effektestimat.

I en metaanalyse ønsker man å estimere δ. Seleksjonsmodellen kan forårsake bias i estimatet for δ. Begg og Mazumdar [8] estimerer den underliggende effekten ved å bruke et vektet

gjennomsnittet, $\bar{t}$, definert i Seksjon 3.2. Den estimerte effekten er beregnet etter seleksjonen. Biasen, β, defineres ved $\beta = \mathrm{E}(\bar{t}) - \delta$.

3.4.2 Parametervalg

For å undersøke hvordan faktorene, nevnt i Seksjon 3.3, påvirker testens styrke, utføres simuleringer med ulike konfigurasjoner av disse faktorene. Parametervalgene er stort sett gjort på bakgrunn av litterære søk. Forfatterne velger to verdier for k, antall studier i metaanalysen. Innen fagfelt som medisin og epidemiologi er $k = 25$ et representativt antall studier. Innen psykologi er $k = 75$ et rimelig valg.

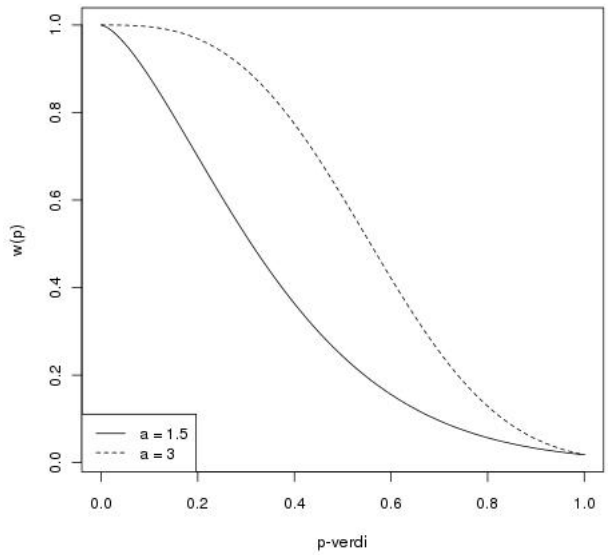
I hver simulering genereres studiene på en slik måte at studiene i metaanalysen, etter publiseringsseleksjon, er delt inn i tre grupper. Hver gruppe har omtrent den samme variansen og består av cirka like mange studier. For $k = 25$ vil den midterste gruppen inkludere én studie mer enn de to andre gruppene. To typer variansspredning brukes: stor ($v = 0.1, 1.0, 10.0$) og liten ($v = 0.5, 1.0, 2.0$). Disse valgene representerer variansspredning på henholdsvis 2.0 og 0.6 på en logaritmisk skala med base 10. Notasjonen $v = 0.1, 1.0, 10.0$ og $v = 0.5, 1.0, 2.0$ beskriver stor og liten variansspredning henholdsvis. Hvis ikke annet er oppgitt, skal en forstå notasjonen som at hver gruppe inneholder omtrent like mange studier. Eksempelvis betyr kombinasjonen $v = 0.5, 1.0, 2.0$ og $k = 25$ at 8 studier har varians 0.5, 9 studier har varians 1.0 og 8 studier har varians 2.0. For å unngå ties, vil variansene innad i en gruppe varieres noe. Dette utdypes nærmere i neste seksjon.

I nulltilfellet er den sanne effektstørrelsen lik null. Den underliggende effektparameteren, δ, varieres fra null til tre standardavvik fra nullverdien. Ett standardavvik settes lik 1.0. Dette er et rimelig valg fordi det representerer standardavviket til studiene i den midterste gruppen.

Konstantene a og b i vektfunksjonene kan varieres for å oppnå større eller mindre grad av seleksjonsjonsbias. Forfatterne setter $a = 1.5$ og $b = 4$ når de ønsker stor grad av seleksjonsjonsbias. De velger $a = 3$ og $b = 4$ når de vil demonstrere moderat seleksjonsbias. Sannsynligheten for at en studie publiseres avtar med økende p-verdi, som vist i Figur 2. Metaanalyser uten publikasjonsbias genereres ved å sette vektfunksjonen lik en.

3.4.3 Utføring

Kort fortalt estimeres den empiriske styrken og det empiriske nivået ved å utføre Monte Carlo-eksperimenter, i tråd med algoritmene gitt av Rizzo [46]. Metaanalysene simuleres. Vi har gitt k varianser, $v_1, v_2, \ldots, v_k$. Ett effektestimat trekkes tilfeldig fra en normalfordeling med forventning δ og varians v_1. Verdien til den valgte vektfunksjonen, w, beregnes. Denne gir sannsynligheten for om studien publiseres. Om studien publiseres eller ikke, avgjøres ved



Figur 2: Seleksjonsmekanisme.

å trekke tilfeldig fra en bernoullifordeling med parameter w. Prosedyren, fra en genererer effektestimatet til en konkluderer om studien skal publiseres eller ikke, gjentas til en studie med varians v_1 inkluderes i metaanalysen. Deretter genereres et nytt effektestimat, denne gangen med varians v_2. Prosessen repeteres til vi har k publiserte studier med de ønskede variansene, $v_1, v_2, \ldots, v_k$.

For å unngå ties, må en unngå like varianser og like effektstørrelser. Det forklares at dette kan gjøres ved å forandre variansen ved hvert nivå med en vilkårlig liten størrelse for hver gang en studie inkluderes i metaanalysen. Simuleringer viser at hvordan en velger disse størrelsene har liten betydning for resultatet. Jeg velger å tillegge en liten, fast $\epsilon = 0.0001$. De estimerte effektstørrelsene genereres fra normalfordelinger. Normalfordelingen er kontinuerlig. Sannsynligheten for at to effektestimater er like er lik null.

Rangkorrelasjonstesten utføres. Et tosidig nominelt signifikansnivå på 0.05 velges for å teste nullhypotesen om ingen publikasjonsbias. Den underliggende effekten estimeres ved $\bar{t}$. Vi beregner hvor mange studier som må til for å generere k publiserte studier.

Prosedyren jeg nå har beskrevet gjentas 5000 ganger. Ut fra disse resultatene beregnes det empiriske nivået eller den empiriske styrken, biasen i estimatet for δ og andelen av de genererte studiene som inkluderes i metaanalysene.

Vedlegg C gir eksempel på simuleringskode. Alle simuleringer i denne oppgaven utføres i statistikkprogrammet R [43]. Resultatene presenteres i Kapittel 4.

3.5 Hypoteser forbundet med testen for publikasjonsbias

Begg og Mazumdar [8] hevder at deres test for publikasjonsbias i realiteten er en test hvor nullhypotesen kan defineres ved

$$H_0 : \beta = 0.$$

Biasen, β, er definert i Seksjon 3.4.1. Samtidig må en ikke glemme at nullhypotesen evaluert ved Kendalls tau uttrykkes ved

$$H_0 : t_i^* \text{ og } v_i \text{ er uavhengige}, \quad i = 1, 2, \ldots, k.$$

Når en skal vurdere testens egenskaper i forhold til å avdekke eventuell publikasjonsbias, er det testens styrke som må vektlegges. Skal man derimot vurdere den praktiske betydningen av publikasjonsbias, må man også ta hensyn til biasen, β. Dersom seleksjonsmodellen forårsaker vesentlig bias i estimatet for δ, er det viktig at testen for publikasjonsbias gir god styrke. I dette tilfellet vil publikasjonsbias påvirke den estimerte effekten, $\bar{t}$, i betydelig grad. Hvis skjevheten i estimatet for δ ikke er vesentlig, vil dårlig styrke være et mindre problem. Lav styrke betyr at testen presterer dårlig med hensyn på å identifisere publikasjonsbias. Den praktiske betydningen av ikke å oppdage publikasjonsbias vil derimot være liten.

4 Simuleringsresultater for Begg og Mazumdars testmetode

4.1 Simuleringsresultater for metaanalyser med publikasjonsbias

4.1.1 Kontroll av Begg og Mazumdars simuleringsresultater

Egenskapene til Begg og Mazumdars testmetode undersøkes. Tabell 1-6 gir simuleringsresultater for de genererte metaanalysene med publikasjonsbias. Resultatene er basert på simuleringsprosedyren og scenarioene beskrevet i Kapittel 3. De to første tabellene viser styrke, prosent inkluderte studier og bias for ensidig seleksjon basert på p-verdi. Tabell 1 viser resultater for små metaanalyser, hvor $k = 25$, mens Tabell 2 gir tilsvarende resultater for store metaanalyser, hvor $k = 75$. Tabell 3 og 4 viser resultater for ensidig seleksjon basert på effektstørrelse for henholdsvis små og store metaanalyser. De to siste tabellene gir resultater for tosidig seleksjon basert på p-verdi. Tabell 5 viser til resultater for $k = 25$, Tabell 6 gir resultater for $k = 75$. Jeg ønsker å kontrollere resultatene til Begg og Mazumdar, som ser på de samme situasjonene. Jeg sammenlikner simuleringsresultatene mine med artikkelens resultater [8].

Det er nødvendig å referere til estimater i tabellene. Jeg henviser til estimatene ved å referere til plasseringer i en matrise. Estimatene tenkes plassert i en 7×4-matrise. Eksempelvis viser plassering (1,1) til styrke, prosent inkluderte studier og bias estimert ved sterk seleksjonsstyrke, stor variansspredning og $\delta = 0$.

Styrken er beregnet ut fra den relative frekvensen hvor rangkorrelasjonstesten er nominelt signifikant ved et tosidig nivå på 0.05. Dersom en forkaster nullhypotesen om ingen publikasjonsbias, anses dette som en suksess. Ikke-forkastning regnes som en fiasko. Metaanalysene genereres uavhengig av hverandre. Forsøkene, hver gang vi beslutter å forkaste eller ikke å forkaste nullhypotesen, er også uavhengige. Sannsynligheten for suksess i hvert forsøk er konstant. Jeg kan bruke en binomisk modell for å beregne standardfeil for styrkeestimatene.

Styrkeestimatene har en standardfeil som kan uttrykkes ved hjelp av

$$\mathrm{SE} = \sqrt{(y/n)\big(1-(y/n)\big)/n},$$

se Vedlegg A. Her er n antall uavhengige forsøk, og y er den observerte verdien av Y, antall suksesser. For å finne den største standardfeilen styrkeestimatene kan ha, deriveres SE med hensyn på y. Dette uttrykket settes lik null slik at

$$\frac{d}{dy}\mathrm{SE} = \frac{1}{2}\frac{1}{\sqrt{\dfrac{(y/n)\big(1-(y/n)\big)}{n}}}\frac{1}{n^2}\big(1-(2y/n)\big) = 0.$$

Jeg løser for y og får $y = n/2$. Siden $\dfrac{d^2\mathrm{SE}}{dy^2} < 0$ når $y = n/2$, gir andrederiverttesten at vi har et lokalt maksimum når $y = n/2$. Videre funksjonsdrøfting viser at SE har et absolutt maksimum i dette punktet.

Tabell 1: Styrke ved ensidig seleksjon basert på p-verdi. Liten metaanalyse ($k = 25$).

	Styrke [% inkluderte studier, bias]			
Seleksjonsstyrke	Sterk		Moderat	
Variansspredning	Stor	Liten	Stor	Liten
Behandlingseffekt (δ)				
.0	61%	23%	35%	15%
	[36%, .34]	[37%, .74]	[57%, .25]	[57%, .54]
.5	54%	21%	25%	12%
	[54%, .16]	[52%, .54]	[74%, .09]	[73%, .35]
1.0	40%	18%	14%	10%
	[64%, .07]	[67%, .36]	[82%, .04]	[85%, .20]
1.5	30%	14%	9%	7%
	[72%, .05]	[79%, .23]	[87%, .02]	[92%, .11]
2.0	20%	9%	6%	5%
	[78%, .03]	[88%, .13]	[90%, .01]	[96%, .05]
2.5	14%	6%	5%	4%
	[82%, .02]	[93%, .08]	[92%, .01]	[98%, .02]
3.0	10%	5%	3%	4%
	[85%, .02]	[96%, .04]	[94%, .01]	[99%, .01]

Tabell 2: Styrke ved ensidig seleksjon basert på p-verdi. Stor metaanalyse ($k = 75$).

	Styrke [% inkluderte studier, bias]			
Seleksjonsstyrke	Sterk		Moderat	
Variansspredning	Stor	Liten	Stor	Liten
Behandlingseffekt (δ)				
.0	99%	60%	88%	37%
	[36%, .34]	[36%, .74]	[56%, .25]	[56%, .54]
.5	98%	60%	77%	31%
	[53%, .16]	[52%, .54]	[74%, .09]	[73%, .34]
1.0	94%	51%	54%	22%
	[64%, .07]	[67%, .36]	[82%, .04]	[85%, .19]
1.5	85%	38%	35%	13%
	[71%, .04]	[79%, .23]	[86%, .02]	[92%, .10]
2.0	71%	23%	20%	7%
	[77%, .03]	[88%, .13]	[90%, .02]	[96%, .05]
2.5	55%	13%	14%	5%
	[81%, .02]	[93%, .08]	[92%, .01]	[98%, .02]
3.0	38%	8%	8%	5%
	[85%, .02]	[96%, .04]	[94%, .01]	[99%, .01]

Tabell 3: Styrke ved ensidig seleksjon basert på effektstørrelse. Liten metaanalyse ($k = 25$).

	Styrke [% inkluderte studier, bias]			
Seleksjonsstyrke	Sterk		Moderat	
Variansspredning	Stor	Liten	Stor	Liten
Behandlingseffekt (δ)				
.0	57%	26%	47%	20%
	[35%, .24]	[36%, .72]	[57%, .17]	[57%, .52]
.5	47%	20%	32%	15%
	[52%, .18]	[52%, .54]	[72%, .10]	[73%, .34]
1.0	35%	14%	20%	10%
	[66%, .11]	[68%, .38]	[81%, .05]	[85%, .21]
1.5	22%	10%	10%	7%
	[77%, .07]	[80%, .24]	[86%, .03]	[92%, .11]
2.0	14%	7%	7%	5%
	[84%, .04]	[89%, .14]	[90%, .02]	[96%, .06]
2.5	8%	5%	5%	4%
	[89%, .02]	[95%, .07]	[93%, .01]	[98%, .03]
3.0	5%	5%	3%	4%
	[91%, .01]	[97%, .04]	[94%, .01]	[99%, .01]

Tabell 4: Styrke ved ensidig seleksjon basert på effektstørrelse. Stor metaanalyse ($k = 75$).

	Styrke [% inkluderte studier, bias]			
Seleksjonsstyrke	Sterk		Moderat	
Variansspredning	Stor	Liten	Stor	Liten
Behandlingseffekt (δ)				
.0	99%	71%	98%	55%
	[34%, .24]	[36%, .72]	[56%, .17]	[56%, .52]
.5	98%	56%	91%	39%
	[51%, .17]	[51%, .54]	[71%, .10]	[72%, .34]
1.0	93%	39%	75%	24%
	[66%, .11]	[67%, .38]	[80%, .05]	[84%, .20]
1.5	78%	25%	50%	11%
	[77%, .07]	[80%, .24]	[86%, .03]	[92%, .11]
2.0	55%	14%	29%	6%
	[84%, .04]	[89%, .14]	[90%, .02]	[96%, .05]
2.5	34%	8%	17%	5%
	[88%, .02]	[94%, .07]	[92%, .01]	[98%, .02]
3.0	19%	5%	10%	4%
	[91%, .01]	[97%, .04]	[94%, .01]	[99%, .01]

Tabell 5: Styrke ved tosidig seleksjon basert på p-verdi. Liten metaanalyse ($k = 25$).

	Styrke [% inkluderte studier, bias]			
Seleksjonsstyrke	Sterk		Moderat	
Varianspredning	Stor	Liten	Stor	Liten
Behandlingseffekt (δ)				
.0	2%	5%	2%	5%
	[36%, .00]	[37%, -.01]	[57%, -.00]	[57%, .00]
.5	12%	9%	7%	7%
	[46%, .17]	[42%, .41]	[66%, .10]	[62%, .25]
1.0	23%	21%	11%	12%
	[54%, .08]	[53%, .46]	[73%, .05]	[72%, .27]
1.5	27%	26%	11%	14%
	[61%, .06]	[67%, .34]	[78%, .03]	[82%, .20]
2.0	28%	19%	12%	9%
	[67%, .04]	[78%, .22]	[82%, .02]	[90%, .12]
2.5	25%	12%	8%	7%
	[72%, .03]	[87%, .13]	[86%, .02]	[95%, .07]
3.0	20%	7%	7%	5%
	[76%, .02]	[93%, .08]	[88%, .01]	[97%, .04]

Tabell 6: Styrke ved tosidig seleksjon basert på p-verdi. Stor metaanalyse ($k = 75$).

	Styrke [% inkluderte studier, bias]			
Seleksjonsstyrke	Sterk		Moderat	
Varianspredning	Stor	Liten	Stor	Liten
Behandlingseffekt (δ)				
.0	2%	4%	2%	5%
	[36%, .00]	[36%, -.00]	[57%, -.00]	[56%, .00]
.5	24%	14%	11%	9%
	[45%, .17]	[41%, .41]	[65%, .10]	[61%, .25]
1.0	54%	53%	24%	25%
	[54%, .08]	[53%, .45]	[73%, .04]	[71%, .27]
1.5	68%	65%	29%	30%
	[60%, .05]	[66%, .34]	[77%, .03]	[82%, .19]
2.0	72%	53%	29%	21%
	[66%, .04]	[78%, .22]	[81%, .02]	[89%, .12]
2.5	71%	32%	28%	11%
	[71%, .03]	[87%, .13]	[85%, .02]	[94%, .07]
3.0	69%	15%	26%	6%
	[75%, .02]	[93%, .08]	[87%, .01]	[97%, .04]

Simuleringsprosedyren gjentas 5000 ganger. Jeg setter $y = n/2$ og $n = 5000$ inn i uttrykket for SE og finner at maksimum standardfeil for styrkeestimatene er tilnærmet lik 0.00707. La p_1 være sannsynligheten for å forkaste nullhypotesen om ingen publikasjonsbias hos Begg og Mazumdar. Likeledes er p_2 sannsynligheten for å forkaste nullhypotesen i mine simuleringer. Kan en anta $p_1 = p_2$? Et tosidig konfidensintervall for $p_1 - p_2$ blir vanligvis brukt for å svare på dette spørsmålet. Avrundingsfeil gjør det derimot vanskelig å kontrollere tabellene i Begg og Mazumdars artikkel basert på statistiske prosedyrer alene. Jeg vil ikke nytte hypotesetesting, men lager en enkel regel for å vurdere rimeligheten av forfatternes estimater. Jeg vil anta at resultatene samsvarer dersom mine styrkeestimater er innen to standardavvik fra resultatene til Begg og Mazumdar [8]. Denne regelen tar dessverre ikke hensyn til at også Begg og Mazumdars estimater avviker fra den virkelige styrken.

Det finnes en mulighet for at et avvik på 1 prosentpoeng i virkeligheten er et avvik på mer enn to standardavvik. På grunn av avrundingsfeil vil det være vanskelig å gjennomføre en rettferdig sammenlikning av resultatene om ikke avvik på denne størrelsen aksepteres.

I Tabell 2, posisjon (2,4), er styrkeestimatet 31%. Dette er avrundet fra 31.42%. Begg og Mazumdar [8] estimerer den tilsvarende verdien til 33%. Tallet 33 er avrundet fra et tall i intervallet [32.5, 33.5). Det er mulig at et avvik på 2 prosentpoeng er innenfor rimelighetens grenser. Avviket kan også være for stort. Jeg forventer noen estimater med dette avviket. Totalt finner jeg 14, ikke medregnet kolonne 1 i Tabell 5.

To styrkeestimater har et avvik på 3 prosentpoeng. Avvik av denne størrelsesordenen krever ekstra oppmerksomhet. Det ene estimatet finnes i Tabell 2, plassering (6,3). Jeg estimerer styrkeestimatet til 14%, mens Begg og Mazumdar [8] estimerer det til 17%. Jeg har gjentatt simuleringene for dette estimatet for å undersøke om avviket skyldes tilfeldigheter hos meg. Det er ikke tilfellet. Avviket kan likevel skyldes tilfeldigheter i simuleringene til Begg og Mazumdar, men også trykkfeil. Det andre estimatet finnes i Tabell 6, plassering (4,2). Mitt styrkeestimat er 65%, mens forfatterne estimerer det til å være 68%.

Jeg foretar en tilsvarende vurdering for prosent inkluderte studier og bias. For disse observatorene er variansen ukjent, men sampelstørrelsen stor ($n = 5000$). Standardfeilen beregnes ved $s/\sqrt{n}$, hvor s er den observerte verdien av sampelstandardavviket S. Legg merke til et stort avvik i prosent inkluderte studier i Tabell 2, posisjon (7,3). 94% av mine genererte studier inkluderes i metaanalysene, i motsetning til 97% av Begg og Mazumdars genererte studier. Legg også merke til avviket i bias i samme tabell, plassering (7,2). Jeg har estimert biasen til 0.04. Forfatterne har estimert den til å være 0.07.

Kolonne 1 i Tabell 5 krever ekstra oppmerksomhet. Jeg klarer ikke å reprodusere Begg og Mazumdars resultater. Når en øker antall studier i en metaanalyse fra $k = 25$ til $k = 75$ uten å forandre øvrige parametre, viser samtlige simuleringsresultater ellers at prosent inkluderte

studier og bias forblir omtrent de samme. Hvorfor skulle dette forandre seg når seleksjonen er tosidig, variansspredningen stor og seleksjonsstyrken sterk? Styrken estimeres generelt for høyt hos Begg og Mazumdar sammenliknet med mine resultater for denne kolonnen. Dette har trolig sammenheng med at en større andel av de genererte studiene inkluderes i mine simulerte metaanalyser. Eksempelvis estimerer Begg og Mazumdar styrken til 48% når $\delta = 1.5$. 27% av de genererte studiene publiseres. Biasen estimeres til 0.12. Jeg estimerer den tilsvarende styrken til 27%, mens 61% av de simulerte studiene inkluderes i metaanalysene. Biasen er 0.06. Jeg velger å se bort fra Begg og Mazumdars [8] resultater i dette tilfellet. Disse er trolig feil.

Ved å bruke en god porsjon skjønn med utgangspunkt i statistiske beregninger, konkluderer jeg at mine simuleringsresultater hovedsaklig samsvarer med resultatene gitt av Begg og Mazumdar [8]. Unntaket er først og fremst den overnevnte kolonnen i Tabell 5. I de få tilfellene hvor avvikene er store, velger jeg å stole på mine resultater. Her har jeg gjentatt simuleringene for å forsikre meg om at avviket ikke skyldes tilfeldigheter eller trykkfeil hos meg. Jeg velger å bruke den vedlagte simuleringskoden som grunnlag for videre arbeid.

4.1.2 Vurdering av testmetodens egenskaper

Det er viktig å vurdere testmetodens egenskaper og undersøke hvordan ulike faktorer påvirker testens styrke. Begg og Mazumdars artikkel [8] behandler dette grundig. Jeg ser ikke behov for en like omfattende gjennomgang i denne oppgaven. Jeg vil kort presentere de viktigste funnene fra mine simuleringsresultater.

Når en skal vurdere tabellene, er det viktig å huske at den underliggende effektparameteren, δ, er uttrykt i standardavvikenheter relativt til variansen til effektestimatet i den gjennomsnittlige studien ($v = 1.0$). Testens egenskaper avhenger av konfigurasjonen $\delta/v^{1/2}$, og ikke av effektstørrelsens absoluttverdi [8].

Først vurderer jeg simuleringsresultatene i Tabell 1 og Tabell 2. Her er seleksjonen ensidig og seleksjonsmodellen avhenger av p-verdien for hypotesen om at den underliggende effekten er lik null. Testen oppnår best styrke når seleksjonsstyrken er høy og variansspredningen stor, både for $k = 25$ og $k = 75$. I tråd med teori om hypotesetesting oppnås bedre styrke ved å øke sampelstørrelsen. Testen har generelt lav styrke når metaanalysen inneholder få studier. Styrken er nokså god når metaanalysen inneholder 75 studier og den underliggende effektparameteren ikke ligger for mange standardavvik fra nullverdien. Unntaket er hovedsaklig ved moderat seleksjonsstyrke i kombinasjon med liten variansspredning.

Tabellene gir også informasjon om den gjennomsnittlige andelen av studier som inkluderes i en metaanalyse, samt biasen, β. Når en skal vurdere simuleringsresultatene, bør en ikke bare se på styrken, men også de to andre observatorene. Styrken synker jo lenger den

underliggende effekten beveger seg bort fra nullverdien. Dette er ikke overraskende. En større andel av de genererte studiene inkluderes i metaanalysen når δ øker [8]. Problemet knyttet til publikasjonsbias minker [35]. Det er betryggende at testen oppnår best styrke når δ er relativt nær nullverdien. Fravær av styrke er et mindre problem for store verdier av δ, da biasen i $\bar{t}$, forårsaket av selektiv publikasjon, er relativt liten [8].

I Tabell 3 og Tabell 4 avhenger seleksjonsmodellen av det observerte effektestimatet. Seleksjonen er ensidig. Tabellene viser de samme tendensene som Tabell 1 og 2. Både seleksjonsstyrke, variansspredning og den underliggende effektparameteren virker inn på styrke, andel studier som inkluderes i metaanalysene og bias. Ved moderat seleksjonsstyrke er testmetodens styrke hovedsaklig noe høyere for seleksjon basert på det observerte effektestimatet enn for en seleksjonsmodell som avhenger av p-verdien. Ved sterk seleksjonsstyrke er det stort sett omvendt, med noen unntak. Her kan en likevel undres om sammenlikningen er reell. Vil seleksjonsstyrken være lik for gitte a og b for de ulike seleksjonsmodellene?

Testmetoden til Begg og Mazumdar er ikke egentlig designet for situasjonen med tosidig seleksjon. Disse simuleringsresultatene vises i Tabell 5 og Tabell 6. Når $\delta = 0$, vil de genererte studiene være symmetrisk fordelt om null. Enhver seleksjonsfunksjon som avhenger av en tosidig p-verdi vil resultere i et funnelplott hvor symmetrien er opprettholdt og korrelasjonen mellom t og v er null [8]. Styrken er lavere enn det nominelle signifikansnivået. På bakgrunn av dette velger jeg å se bort fra tosidig seleksjon i den videre oppgaven, til tross for minimal bias når $\delta = 0$. Testen gir bedre styrke for verdier av δ mellom 1.0 og 2.5. Seleksjonspress forårsaker i disse tilfellene korrelasjon mellom t og v.

Simuleringsresultatene bekrefter at testens styrke varierer og avhenger av ulike faktorer. Antall studier i metaanalysen og variansspredningen er kjente størrelser, mens seleksjonsmekanismen og den sanne effekten er ukjente faktorer. Generelt oppnår ikke testmetoden til Begg og Mazumdar god styrke. Styrken er dårlig når δ ligger langt fra nullverdien. Den er heller ikke god når metaanalysen inneholder få studier, selv ikke når seleksjonsstyrken er sterk. Selv om en ikke forkaster nullhypotesen om ingen publikasjonsbias, bør en ikke utelukke selektiv publikasjon.

4.2 Simuleringsresultater for metaanalyser uten publikasjonsbias

Tabell 7 og Tabell 8 viser simuleringsresultater for metaanalyser uten publikasjonsbias. Den første tabellen gir resultater for små metaanalyser. Den andre presenterer resultater for store metaanalyser. Begg og Mazumdar [8] inkluderer ikke tabeller for disse situasjonene, men kommenterer kort at det "nominelle signifikansnivået" er lavere enn 5% i alle tilfeller. Her mistenker jeg at forfatterne egentlig ønsker å si at det virkelige signifikansnivået er lavere enn det nominelle signifikansnivået på 5%.

Tabell 7: Nivå. Liten metaanalyse ($k = 25$).

	Nivå [% inkluderte studier, bias]	
Varianspredning	Stor	Liten
Behandlingseffekt (δ)		
.0	1,94%	4,42%
	[100%, .00]	[100%, -.00]
.5	1,90%	4,36%
	[100%, .00]	[100%, -.00]
1.0	2,28%	4,18%
	[100%, .00]	[100%, -.00]
1.5	2,16%	4,30%
	[100%, .00]	[100%, -.00]
2.0	2,08%	4,56%
	[100%, .00]	[100%, -.01]
2.5	2,00%	4,42%
	[100%, -.00]	[100%, -.00]
3.0	2,18%	4,46%
	[100%, .00]	[100%, .00]

Tabell 8: Nivå. Stor metaanalyse ($k = 75$).

	Nivå [% inkluderte studier, bias]	
Varianspredning	Stor	Liten
Behandlingseffekt (δ)		
.0	1,88%	4,54%
	[100%, -.00]	[100%, -.00]
.5	2,16%	4,40%
	[100%, .00]	[100%, -.00]
1.0	2,20%	4,72%
	[100%, -.00]	[100%, .00]
1.5	1,86%	4,34%
	[100%, -.00]	[100%, .00]
2.0	1,64%	4,20%
	[100%, .00]	[100%, .01]
2.5	1,94%	4,58%
	[100%, -.00]	[100%, -.00]
3.0	1,86%	4,46%
	[100%, .00]	[100%, .00]

Signifikansnivået estimeres med to ekstra desimaler sammenliknet med styrken i de fire første tabellene. Jeg ser behov for et høyere presisjonsnivå når jeg skal vurdere nivået. Kan det virkelige signifikansnivået antas å være lik det nominelle?

Jeg ser på tilfellene $k = 25$ og $k = 75$ samlet. Ved stor variansspredning er nivåestimatene spredt fra 0.0164 til 0.0228. Tosidige konfidensintervaller for det virkelige nivået, beregnet ut fra disse to ytterpunktene, er gitt ved henholdsvis [0.01288, 0.01992] og [0.01866, 0.02694]. Konfidensnivået er 0.95. De andre estimatene vil naturligvis gi konfidensintervaller med ytterpunkter en plass mellom 0.01288 og 0.02694. Vi forkaster hypotesen om at det virkelige nivået er lik det nominelle for samtlige tilfeller.

Ved liten variansspredning er nivåestimatene høyere. Det laveste nivået er estimert til 0.0418, mens det høyeste er 0.0472. Tosidige konfidensintervaller med konfidensnivå 0.95 er henholdsvis [0.0363, 0.0473] og [0.0413, 0.0531]. Alle de andre estimatene vil gi konfidensintervaller med ytterpunkter mellom 0.0363 og 0.0531. Nullhypotesen om at det virkelige signifikansnivået er lik 0.05 forkastes i åtte av 14 tilfeller.

Verdien av δ ser ikke ut til å være en faktor som virker inn på testmetodens nivå når effektestimatene er normalfordelte.

4.3 Reell variansfordeling

I Begg og Mazumdars testsituasjon er studiene i metaanalysene inndelt i tre grupper. Gruppeinndelingen er basert på effektestimatenes varianser. Hver gruppe inneholder, så godt det lar seg gjøre, like mange studier. Denne variansfordelingen er ikke typisk for reelle metaanalyser. Vil variansfordelingen være en faktor som virker inn på testmetodens nivå og styrke?

Enkle simuleringer bekrefter dette. Jeg velger $k = 25$ og inkluderer 14 studier med varians 10.0 og 9 studier med varians 1.0. De resterende to studiene har varians 0.1. Videre lar jeg den underliggende effekten, δ, være lik null. Simuleringsprosedyren beskrevet i Kapittel 3 repeteres 5000 ganger. Stadig settes det tosidige nominelle nivået lik 0.05. I denne situasjonen estimeres det tosidige nivået til 0.0274. Det tilsvarende nivåeestimatet er 0.0194 med Begg og Mazumdars variansfordeling. La p_1 og p_2 være sannsynligheten for suksess med henholdsvis den nye og Begg og Mazumdars variansfordeling. Et tosidig konfidensintervall for $p_1 - p_2$ med konfidensnivå 0.95 viser at variansfordelingen vil påvirke testmetodens nivå. Med sterk seleksjonsstyrke og ensidig seleksjon basert på p-verdi, estimeres styrken til 39% med den kanskje mer reelle variansfordelingen. Dette er lavere enn det tilsvarende estimatet med variansfordelingen til Begg og Mazumdar, hvor styrkeestimatet er 61%.

Jeg ønsker ikke en inngående diskusjon rundt denne tematikken på det nåværende tidspunktet. Det registreres et behov for å studere testmetodens styrke dersom vi har en rimelig fordeling av variansene. Simuleringer i Kapittel 11 vil dekke dette ønsket.

5 Mulige årsaker til det feilaktige signifikansnivået for testmetoden introdusert av Begg og Mazumdar

Resultatene viser at det virkelige signifikansnivået hovedsaklig ikke kan antas å være lik det nominelle for Begg og Mazumdars testmetode [8]. I dette kapittelet undersøkes årsakene til det feilaktige nivået.

Første del av kapittelet prøver å svare på om det dårlige nivået skyldes problemer knyttet til testen basert på Kendalls tau generelt, eller om problemet tilhører den spesielle situasjonen innført av Begg og Mazumdar [8]. Enkle simuleringer gjennomføres.

Den andre delen av kapittelet undersøker om Begg og Mazumdar innfører en testsituasjon som bryter med de vanlige forutsetningene for å utføre en rangkorrelasjonstest under nullhypotesen. Denne biten er mer teoretisk.

5.1 Signifikansnivået til rangkorrelasjonstesten basert på Kendalls tau når alle forutsetninger er oppfylt

Jeg undersøker nivået til rangkorrelasjonstesten basert på Kendalls tau i tilfeller hvor jeg vet at alle forutsetningene er oppfylt under nullhypotesen. Eksempelvis trekker jeg k iid X-variable fra en standardnormalfordeling. Likeledes trekker jeg k iid Y-variable fra den samme fordelingen. Jeg korrelerer X og Y ved hjelp av Kendalls tau. Deretter trekkes k iid X-variable fra en uniform fordeling på intervallet $[1, 10]$. Disse variablene korreleres med k iid Y variable fra en Gamma(0.1,1)-fordeling. De stokastiske variablene X og Y er uavhengige i alle tilfeller. Til sammenlikning kjører jeg også rangkorrelasjonstesten basert på Spearmans rho og Pearsons korrelasjonstest. Simuleringene kjøres for $k = 25$ og $k = 75$. Hver simuleringprosedyre repeteres 100000 ganger. Tabell 9 viser en oversikt over de estimerte nivåene i hver situasjon. Det empiriske signifikansnivået er beregnet ved et tosidig nominelt nivå på 0.05.

Tabell 9: Estimert nivå

X	Y	k	Kendalls tau	Spearmans rho	Pearsons korrelasjon
N(0,1)	N(0,1)	25	0.04611	0.04943	0.05006
N(0,1)	N(0,1)	75	0.05027	0.04959	0.04997
Uniform(1,10)	Gamma(0.1,1)	25	0.04652	0.04962	0.03603
Uniform(1,10)	Gamma(0.1,1)	75	0.04983	0.04977	0.04179

Resultatene er basert på funksjonen `cor.test()` i statistikkprogrammet R [43]. Den logiske indikatoren `exact` settes lik `NULL`. Dette er defaultverdien i R. Defaultverdien beregner en eksakt p-verdi for Kendalls tau under nullhypotesen om ingen publikasjonsbias for $k < 50$. For $k \geq 50$ nyttes den asymptotiske normalfordelingen. Simuleringsresultatene i Tabell 15,

Kapittel 7, bekrefter at nivåproblemene eksisterer selv om en nytter den eksakte fordelingen til Kendalls tau. Jeg vil benytte funksjonen `cor.test()` til å beregne resultater i Kapittel 5-8, så sant ikke annet er oppgitt. For mer teori knyttet til den eksakte fordelingen til Kendalls tau, se Kendall og Gibbons [27].

For Kendalls tau forkastes nullhypotesen om at det virkelige nivået er lik det nominelle når $k = 25$ for begge modellsituasjonene. Konklusjonene er basert på tosidige konfidensintervaller med konfidensnivå 0.95. For $k = 75$ gir konfidensintervallene ikke grunnlag for å forkaste den samme hypotesen.

Observatoren $S = C - D$ kan bare ta et endelig antall verdier. Den eksakte fordelingen til S er diskret. Det kan være vanskelig å oppnå et korrekt nivå, det vil si et virkelig nivå som samsvarer med det nominelle. Problemet avtar med økende k.

Ingen konfidensintervaller med konfidensnivå 0.95 gir grunnlag til å forkaste hypotesen om at nivået er lik 0.05 for testen basert på Spearmans rho. Den eksakte fordelingen til Spearmans rho er også diskret. Observatoren kan derimot ta flere verdier sammenliknet med Kendalls tau. Dette kan forklare hvorfor rangkorrelasjonstesten basert på rho oppnår korrekt nivå for lavere verdier av k enn testen basert på tau.

Når variablene er trukket fra normalfordelinger, gir Pearsons korrelasjonstest et virkelig nivå som samsvarer med det nominelle. Forutsetningen om at X og Y har en simultan bivariat normalfordeling brytes når X er uniformt fordelt og Y er trukket fra en gammafordeling. Det reelle nivået er ikke lik det nominelle, vist ved tosidige konfidensintervaller med konfidensnivå 0.95.

Jeg ønsker å undersøke Kendalls tau litt nærmere. Jeg gjentar simuleringene beskrevet ovenfor for $k = 25$ når X og Y er standardnormalfordelte. Denne gangen bruker jeg den asymptotiske fordelingen til Kendalls tau ved beregning av p-verdi. Nivået estimeres til 0.05214. Jeg nytter også den asymptotiske fordelingen til Kendalls tau med kontinuitetskorreksjon. Leseren henvises til Kendall og Gibbons [27] for teori om denne. Nivået estimeres til 0.04625. Nullhypotesen om at nivået er lik 0.05 forkastes i begge tilfeller. Kontinuitetskorreksjonen gir en bedre tilnærming til den eksakte fordelingen til tau i denne situasjonen. Jeg undres hvorfor ikke Begg og Mazumdar har nyttet kontinuitetskorreksjon for $k = 25$.

Simuleringene ovenfor gjentas også når X og Y er standardnormalfordelte og $k = 75$. Jeg beregner p-verdien først ved hjelp av den asymptotiske fordelingen til tau med kontinuitetskorreksjon, deretter beregnes en eksakt p-verdi for hypotesen om ingen publikasjonsbias. Nivået estimeres til henholdsvis 0.04939 og 0.05023. Vi forkaster ikke nullhypotesen om at nivået er lik 0.05 for noen av tilfellene. Betydningen av å nytte kontinuitetskorreksjonen for store verdier av k vil trolig være liten. Jeg holder meg til defaultverdien i R, og vil ikke videre ta hensyn til kontinuitetskorreksjon for $k = 75$ i Kapittel 5-8.

Resultatene i Kapittel 4 kan, for $k = 25$, være påvirket av at vi bruker en diskret test. Dessuten er ikke tilnærmingen mot den asymptotiske normalfordelingen optimal for lave verdier av k. Likevel forklarer ikke dette hvorfor det virkelige nivået i flere tilfeller ser ut til å ligge så lavt som 0.02 for testen introdusert av Begg og Mazumdar. For $k = 75$ viser simuleringene i denne seksjonen at det virkelige nivået ligger tett opp mot det nominelle dersom testens forutsetninger er oppfylt.

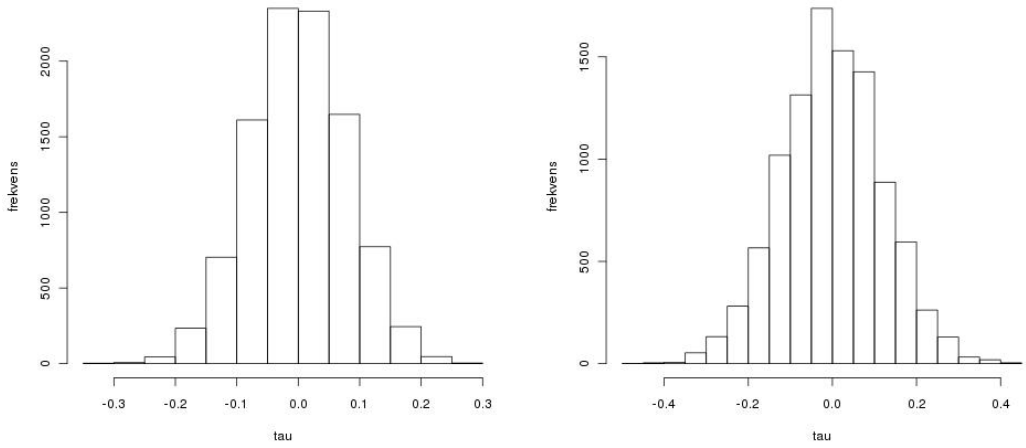
5.2 Sannsynlighetsfordelingen til Kendalls tau i Begg og Mazumdars testmodell

Kan sannsynlighetsfordelingen til Kendalls tau i Begg og Mazumdar testmodell antas å være lik den asymptotiske fordelingen til den ordinære Kendalls tau under nullhypotesen?

La forutsetningene for å utføre en rangkorrelasjonstest basert på Kendalls tau være oppfylt. Under nullhypotesen er Kendalls tau asymptotisk normalfordelt med forventning 0 og varians gitt ved $2(2k + 5)/\big(9k(k - 1)\big)$. Dersom $k = 75$, er variansen til tau lik 0.006206.

Jeg bekrefter den asymptotiske fordelingen til den ordinære Kendalls tau ved hjelp av simuleringer. Jeg trekker 75 iid standardnormalfordelte X-variable og korrelerer dem mot 75 iid standardnormalfordelte Y-variable ved hjelp av Kendalls tau. Alle forutsetningene er tilfredsstilt. Dette repeteres 10000 ganger. Det tosidige nominelle nivået settes lik 0.05. Nivået estimeres til 0.0517. Den estimerte forventningen til tau er 0.001388541, mens variansen estimeres til 0.006254767. Sannsynlighetsfordelingen til tau ser ut til å være normal, basert på visuell inspeksjon av histogrammet for Kendalls tau, Figur 3a. Vi forkaster ikke hypotesen om at forventningen er lik null ved et tosidig signifikansnivå på 0.05. Et 95% tosidig konfidensintervall om den estimerte variansen er [0.006084946, 0.006431832]. Se Vedlegg A for beregning av konfidensintervallet. Vi forkaster ikke hypotesen om at variansen er lik 0.006206.

Jeg undersøker videre sannsynlighetsfordelingen til Kendalls tau i Begg og Mazumdars testmodell. Jeg tar for meg en typisk simuleringssituasjon hvor nivået viser seg ikke å være korrekt. Metaanalysene inneholder $k = 75$ studier og har stor variansspredning. Den underliggende effekten, δ, settes lik null. Simuleringsprosedyren beskrevet i Kapittel 3 gjentas 10000 ganger. Nivået estimeres til 0.0209 ved et tosidig nominelt nivå på 0.05. Det er ikke urimelig å anta at Kendalls tau er normalfordelt, basert på en subjektiv inspeksjon av histogrammet for tau, Figur 3b. Forventningen estimeres til 0.000611027. En forkaster ikke hypotesen om at forventningen er lik null. Variansestimatet er 0.004339671. Et 95% tosidig konfidensintervall for variansen er [0.004221846, 0.004462522]. Hypotesen om at variansen er lik 0.006206 forkastes. Sannsynlighetsfordelingen til Kendalls tau i Begg og Mazumdars testmodell er ikke lik sannsynlighetsfordelingen til den ordinære Kendalls tau.



(a) Histogram for Kendalls tau i ordinær situasjon.

(b) Histogram for Kendalls tau for Begg og Mazumdars testmodell.

Figur 3: Histogram for tau.

5.3 Spearmans rho

Tilsvarende simuleringer som dem presentert i Kapittel 3, utføres også for en rangkorrelasjonstest basert på Spearmans rho. Jeg korrelerer fortsatt de standardiserte effektestimatene mot variansene. Signifikansnivået til testmetoden basert på Spearmans rho estimeres i situasjoner hvor metaanalysene ikke er utsatt for selektiv publikasjon. Hypotesen om at det virkelige signifikansnivået er lik det nominelle forkastes for samtlige verdier av δ når variansspredningen er stor. Også her er nivået lavere enn det nominelle. Nivåestimatene i kombinasjon med styrkeestimatene understøtter Sheskins konklusjon [48]; Kendalls tau og Spearmans rho inneholder mye av den samme informasjonen. Simuleringsresultatene for Spearmans rho inkluderes ikke i oppgaven grunnet tabellenes omfang. Jeg vil i den videre oppgaven se bort fra rangkorrelasjonstesten basert på Spearmans rho. Valg av observator vil trolig ikke spille en avgjørende rolle.

Problemet med det feilaktige nivået ser ikke ut til å være knyttet til en rangkorrelasjonstest basert på Kendalls tau generelt, men den testmodellen Begg og Mazumdar innfører. Simuleringsresultatene i dette kapittelet understøtter denne konklusjonen. Videre undersøkes årsakene til det feilaktige nivået på et mer matematisk og teoretisk plan. Først er det nødvendig med noen innledende antakelser og forklaringer angående notasjon.

5.4 Antakelser

Effektestimatene genereres gitt variansene. Begg og Mazumdar lar formelt sett variansene være faste konstanter i deres simuleringer. Dette kan forenklet sammenliknes med en som velger ut ønskede varianser før innsamling av data. Denne situasjonen er ikke reell.

Når en skal utføre metaanalyser, vil man i realiteten finne fram til relevante studier om et gitt tema. På forhånd vet man ikke studienes effektstørrelse eller varians. Effektestimatene og variansene er observerte verdier av de stokastiske variablene $t_1, t_2, \ldots, t_k$ og $v_1, v_2, \ldots, v_k$.

Det er en konvensjonell oppfatning at de k stokastiske parene $(t_1, v_1), (t_2, v_2), \ldots, (t_k, v_k)$ fra studiene i en metaanalyse er uavhengige og har den samme kontinuerlige bivariate fordelingen. Utledninger og bevis i dette kapittelet forutsetter dette. Det følger at v_i er uavhengig av v_j. Videre er t_i gitt $v_1, v_2, \ldots, v_k$ uavhengig av t_j gitt $v_1, v_2, \ldots, v_k$. Den betingede fordelingen til t_i gitt $v_1, v_2, \ldots, v_k$ er lik den betingede fordelingen til t_i gitt v_i.

5.5 Misbruk av notasjon

Dette kapittelet inneholder noe misbruk av notasjon. Jeg ønsker å holde meg nær notasjonen til Begg og Mazumdar [8]. Dette er ikke problemfritt. Vanligvis vil t_i være den observerte verdien av den stokastiske variabelen T_i. På samme måte vil v_i være den observerte verdien av V_i. Eksempelvis kan en da snakke om den betingede fordelingen til T_i gitt $V_i = v_i$. Den betingede forventningen til T_i gitt $V_i = v_i$ kan uttrykkes ved notasjonen $\mathrm{E}(T_i|V_i = v_i)$ eller $\mathrm{E}(T_i|v_i)$.

De k parene $(t_1, v_1), (t_2, v_2), \ldots, (t_k, v_k)$ er stokastiske. Jeg nytter notasjonen $\mathrm{E}(t_i|v_i)$ for å referere til den betingede forventningen til t_i gitt den observerte verdien av den stokastiske variabelen v_i. Dette til tross for dårlig notasjon. Det bør være klart ut fra de ulike situasjonene om v_i betegner en stokastisk variabel eller den observerte verdien av en stokastisk variabel.

Jeg vil også bruke notasjonen $f(t_i)$ som betegnelse for tetthetsfunksjonen til t_i. Her skal t_i selvsagt oppfattes som argumentet, ikke som en stokastisk variabel. Dersom t_i var den observerte verdien av den stokastiske variabelen T_i, ville notasjonen $f_{T_i}(t_i)$ vært passende.

Hvis jeg i den videre oppgaven eksempelvis sier at korrelasjonen mellom t_i^* og v_i er ulik null, skal dette oppfattes som at $\mathrm{Cor}(t_i^*, v_i) \neq 0$. Hvis jeg derimot sier at t^* og v korreleres, mener jeg vanligvis at t_i^* og v_i korreleres ved hjelp av Kendalls tau for $i = 1, 2, \ldots, k$.

5.6 Variansen til $t_i - \bar{t}$ gitt $v_1, v_2, \ldots, v_k$

Fordi variansene er stokastiske variable, er det underforstått at man bør tolke $t_i \sim N(\delta, v_i)$ som at den betingede fordelingen til t_i gitt v_i er normalfordelt med forventning δ og varians v_i. Det er også underforstått at $v_i^* = v_i - (\sum v_j^{-1})^{-1}$ er den betingede variansen til $t_i - \bar{t}$ gitt

$v_1, v_2, \ldots, v_k$. Det er nødvendig å verifisere dette uttrykket. Dersom uttrykket for variansen til $t_i - \bar{t}$ ikke er korrekt, kan dette være en mulig forklaring på det feilaktige nivået til Begg og Mazumdars test.

Grunnleggende regneregler for varians gir

$$\mathrm{Var}(t_i - \bar{t}|v_1, \ldots, v_k) = \mathrm{Var}(t_i|v_1, \ldots, v_k) + \mathrm{Var}(\bar{t}|v_1, \ldots, v_k) - 2\mathrm{Cov}(t_i\ ,\ \bar{t}|v_1, \ldots, v_k).$$

Det følger fra definisjonene ovenfor at

$$\mathrm{Var}(t_i|v_1, \ldots, v_k) = v_i.$$

Videre er

$$\begin{aligned}
\mathrm{Var}(\bar{t}|v_1, \ldots, v_k) &= \mathrm{Var}\left(\frac{\sum v_j^{-1}t_j}{\sum v_j^{-1}}\Big|v_1, \ldots, v_k\right) = \frac{1}{(\sum v_j^{-1})^2}\mathrm{Var}\left(\sum v_j^{-1}t_j\big|v_1, \ldots, v_k\right) \\
&= \frac{1}{(\sum v_j^{-1})^2}\sum v_j^{-2}\mathrm{Var}(t_j\big|v_1, \ldots, v_k) = \frac{1}{(\sum v_j^{-1})^2}\sum v_j^{-1} \\
&= \frac{1}{\sum v_j^{-1}}.
\end{aligned}$$

Det gjenstår å finne $\mathrm{Cov}(t_i\ ,\ \bar{t}|v_1, \ldots, v_k)$. Dette krever noe mer arbeid enn de to første leddene. Jeg nytter regneregler for kovarians og får at

$$\mathrm{Cov}(t_i\ ,\ \bar{t}|v_1, \ldots, v_k) = \mathrm{E}(t_i\bar{t}|v_1, \ldots, v_k) - \mathrm{E}(t_i|v_1, \ldots, v_k)\mathrm{E}(\bar{t}|v_1, \ldots, v_k).$$

Forventningen til t_i følger per definisjon. Vi har

$$\mathrm{E}(t_i|v_1, \ldots, v_k) = \delta.$$

Enkel regning gir

$$\begin{aligned}
\mathrm{E}(\bar{t}|v_1, \ldots, v_k) &= \mathrm{E}\left(\frac{\sum v_j^{-1}t_j}{\sum v_j^{-1}}\Big|v_1, \ldots, v_k\right) = \frac{1}{\sum v_j^{-1}}\mathrm{E}\left(\sum v_j^{-1}t_j\big|v_1, \ldots, v_k\right) \\
&= \frac{1}{\sum v_j^{-1}}\sum\left(v_j^{-1}\mathrm{E}(t_j|v_1, \ldots, v_k)\right) = \delta.
\end{aligned}$$

Fordi (t_i, v_i) er uavhengig av (t_j, v_j), har vi at

$$\begin{aligned}
\mathrm{E}(t_i\bar{t}|v_1,\ldots,v_k) &= \mathrm{E}\left(t_i\frac{\sum v_j^{-1}t_j}{\sum v_j^{-1}}\Big|v_1,\ldots,v_k\right)\\
&= \frac{1}{\sum v_j^{-1}}\mathrm{E}\left(t_i v_i^{-1}t_i + t_i\sum_{j\neq i} v_j^{-1}t_j\big|v_1,\ldots,v_k\right)\\
&= \frac{1}{\sum v_j^{-1}}v_i^{-1}\mathrm{E}(t_i^2|v_1,\ldots,v_k)\\
&\quad + \frac{1}{\sum v_j^{-1}}\mathrm{E}(t_i|v_1,\ldots,v_k)\mathrm{E}\left(\sum_{j\neq i} v_j^{-1}t_j\big|v_1,\ldots,v_k\right)\\
&= \frac{1}{\sum v_j^{-1}}v_i^{-1}\mathrm{E}(t_i^2|v_1,\ldots,v_k) + \frac{1}{\sum v_j^{-1}}\delta\sum_{j\neq i} v_j^{-1}\delta.
\end{aligned}$$

Videre er

$$\mathrm{E}(t_i^2|v_1,\ldots,v_k) = \mathrm{Var}(t_i|v_1,\ldots,v_k) + \mathrm{E}(t_i|v_1,\ldots,v_k)^2 = v_i + \delta^2.$$

Vi setter dette inn i uttrykket for $\mathrm{E}(t_i\bar{t}|v_1,\ldots,v_k)$ og får

$$\begin{aligned}
\mathrm{E}(t_i\bar{t}|v_1,\ldots,v_k) &= \frac{1}{\sum v_j^{-1}}v_i^{-1}(v_i+\delta^2) + \delta^2\frac{1}{\sum v_j^{-1}}\sum_{j\neq i} v_j^{-1}\\
&= \frac{1}{\sum v_j^{-1}}\left(1 + \delta^2 v_i^{-1} + \delta^2\sum_{j\neq i} v_j^{-1}\right)\\
&= \frac{1}{\sum v_j^{-1}}\left(1 + \delta^2\sum v_j^{-1}\right) = \frac{1}{\sum v_j^{-1}} + \delta^2.
\end{aligned}$$

Ved innsetting finner vi at uttrykket for kovariansen er

$$\mathrm{Cov}(t_i, \bar{t}|v_1,\ldots,v_k) = \frac{1}{\sum v_j^{-1}} + \delta^2 - \delta^2 = \frac{1}{\sum v_j^{-1}}.$$

Vi får

$$\mathrm{Var}(t_i - \bar{t}|v_1,\ldots,v_k) = v_i + \frac{1}{\sum v_j^{-1}} - \frac{2}{\sum v_j^{-1}} = v_i - \frac{1}{\sum v_j^{-1}}.$$

Jeg har med dette bekreftet uttrykket for den betingede variansen til $t_i - \bar{t}$ gitt $v_1, v_2, \ldots, v_k$. Feilen ligger ikke her.

5.7 Fordelingen til t_i^* gitt $v_1, v_2, \ldots, v_k$ i Begg og Mazumdars testsituasjon

Et naturlig steg videre er å finne fordelingen til t_i^* gitt $v_1, v_2, \ldots, v_k$. Jeg vil ha bruk for dette uttrykket ved senere anledninger. Fra før vet vi at den betingede fordelingen til t_i gitt $v_1, v_2, \ldots, v_k$ er normal med forventning δ og varians v_i i testsituasjonen til Begg og

Mazumdar [8]. Gitt variansene er t_i^* en lineær kombinasjon av uavhengige og normalfordelte t-er. Følgelig er også t_i^* gitt $v_1, v_2, \ldots, v_k$ normalfordelt. Normalfordelingen er entydig bestemt ved forventning og varians. Forventningen er gitt ved

$$\mathrm{E}(t_i^*|v_1,\ldots,v_k) = \mathrm{E}\left(\frac{t_i-\bar{t}}{(v_i^*)^{1/2}}|v_1,\ldots,v_k\right) = \frac{1}{(v_i^*)^{1/2}}\big(\mathrm{E}(t_i|v_1,\ldots,v_k) - \mathrm{E}(\bar{t}|v_1,\ldots,v_k)\big) = 0.$$

Variansen er

$$\mathrm{Var}(t_i^*|v_1,\ldots,v_k) = \mathrm{Var}\left(\frac{t_i-\bar{t}}{(v_i^*)^{1/2}}|v_1,\ldots,v_k\right) = \frac{1}{v_i^*}\mathrm{Var}(t_i-\bar{t}|v_1,\ldots,v_k) = \frac{v_i^*}{v_i^*} = 1.$$

Det følger at t_i^* gitt $v_1, v_2, \ldots, v_k$ er standardnormalfordelt for $i = 1, 2, \ldots, k$.

5.8 Er t_i^* uavhengig av $v_1, v_2, \ldots, v_k$ under nullhypotesen i testsituasjonen til Begg og Mazumdar?

Jeg vet enda ikke grunnen til det feilaktige nivået og undersøker derfor forutsetningene for å utføre testen basert på Kendalls tau. Er t_i^* uavhengig av $v_1, v_2, \ldots, v_k$ under nullhypotesen?

Fordelingen til t_i^* gitt $v_1, v_2, \ldots, v_k$ avhenger ikke av $v_1, v_2, \ldots, v_k$ som funksjon i Begg og Mazumdars testsituasjon. Det følger at t_i^* er uavhengig av variansene. Dette kan vises formelt.

Per definisjon vet vi at t_i^* er uavhengig av $v_1, v_2, \ldots, v_k$ dersom $f(t_i^*|v_1, \ldots, v_k) = f(t_i^*)$. Vi kan uttrykke $f(t_i^*, v_1, \ldots, v_k)$ ved $f(t_i^*|v_1, \ldots, v_k)f(v_1, \ldots, v_k)$. Ved hjelp av dette uttrykket kan vi finne den marginale sannsynlighetsfordelingen til t_i^*. Vi har at

$$\begin{aligned} f(t_i^*) &= \int_{v_k}\cdots\int_{v_1} f(t_i^*, v_1, \ldots, v_k)\; dv_1 \ldots dv_k \\ &= \int_{v_k}\cdots\int_{v_1} f(t_i^*|v_1, \ldots, v_k)f(v_1, \ldots, v_k)\; dv_1 \ldots dv_k. \end{aligned}$$

Fordi $f(t_i^*|v_1, \ldots, v_k)$ er standardnormalfordelt, får vi at

$$f(t_i^*) = f(t_i^*|v_1, \ldots, v_k)\int_{v_k}\cdots\int_{v_1} f(v_1, \ldots, v_k)\; dv_1 \ldots dv_k = f(t_i^*|v_1, \ldots, v_k).$$

Også t_i^* er standardnormalfordelt i testsituasjonen til Begg og Mazumdar.

5.9 Bivariat fordeling

Videre undersøker jeg om de k stokastiske parene $(t_1^*, v_1), (t_2^*, v_2), \ldots, (t_k^*, v_k)$ har den samme bivariate fordelingen. Fordi t_i^* er uavhengig av $v_1, v_2, \ldots, v_k$ under nullhypotesen om ingen publikasjonsbias i testsituasjonen til Begg og Mazumdar, kan vi uttrykke $f(t_i^*, v_i) = f(t_i^*)f(v_i)$ og $f(t_j^*, v_j) = f(t_j^*)f(v_j)$.

Det er rimelig å anta at v_i og v_j er trukket fra den samme fordelingen dersom variansene er stokastiske variable. Det samme er tilfellet for t_i^* og t_j^*. I den aktuelle situasjonen er både t_i^* og

t_j^* er standardnormalfordelt. Følgelig har (t_i^*, v_i) og (t_j^*, v_j) den samme bivariate fordelingen. Forutsetningen om identisk fordelte par er oppfylt.

I simuleringssituasjonen beskrevet av Begg og Mazumdar [8] er variansene formelt sett faste konstanter. Også faste konstanter har en sannsynlighetsfordeling og kan oppfattes som stokastiske variabler. Punktsannsynligheten for v_i kan beskrives ved

$$f(v_i) = \begin{cases} 1 & \text{hvis } v_i = c \\ 0 & \text{hvis } v_i \neq c, \end{cases}$$

hvor c er den gitte aktuelle verdien for v_i. På samme måte definerer

$$f(v_j) = \begin{cases} 1 & \text{hvis } v_j = d \\ 0 & \text{hvis } v_j \neq d \end{cases}$$

punktsannsynligheten for v_j, hvor d er den gitte aktuelle verdien for v_j. De ulike parene har ikke den samme bivariate fordelingen i denne simuleringsmodellen. Jeg vil senere argumentere for at dette trolig ikke vil skape problemer.

5.10 Uavhengige par

Er $(t_1^*, v_1), (t_2^*, v_2), \ldots, (t_k^*, v_k)$ k uavhengige todimensjonale stokastiske vektorer? Jeg undersøker om t_i^* er uavhengig av t_j^* gitt v_1,v_2,...,v_k og vurderer $\text{Cov}(t_i^*\ ,\ t_j^* | v_1, ..., v_k)$. Vi har at

$$\begin{aligned}
\text{Cov}(t_i^*\ ,\ t_j^* | v_1, ..., v_k) &= \text{Cov}\left(\frac{1}{(v_i^*)^{1/2}} \left(t_i - \frac{\sum v_l^{-1} t_l}{\sum v_l^{-1}} \right)\ ,\ \frac{1}{(v_j^*)^{1/2}} \left(t_j - \frac{\sum v_l^{-1} t_l}{\sum v_l^{-1}} \right) | v_1, ..., v_k \right) \\
&= \frac{1}{(v_i^*)^{1/2}(v_j^*)^{1/2}} \\
&\quad \times \text{Cov}\left(\frac{t_i \sum v_l^{-1} - \sum v_l^{-1} t_l}{\sum v_l^{-1}}\ ,\ \frac{t_j \sum v_l^{-1} - \sum v_l^{-1} t_l}{\sum v_l^{-1}} | v_1, ..., v_k \right) \\
&= \frac{1}{(v_i^*)^{1/2}(v_j^*)^{1/2}(\sum v_l^{-1})^2} \\
&\quad \times \text{Cov}\left(t_i \sum v_l^{-1} - \sum v_l^{-1} t_l\ ,\ t_j \sum v_l^{-1} - \sum v_l^{-1} t_l | v_1, ..., v_k \right).
\end{aligned}$$

Jeg forenkler slik at

$$\begin{aligned}
&\text{Cov}\left(t_i \sum v_l^{-1} - \sum v_l^{-1} t_l\ ,\ t_j \sum v_l^{-1} - \sum v_l^{-1} t_l | v_1, ..., v_k \right) \\
&\quad = \text{Cov}\left(t_i \sum v_l^{-1}\ ,\ t_j \sum v_l^{-1} | v_1, ..., v_k \right) - \text{Cov}\left(t_i \sum v_l^{-1}\ ,\ \sum v_l^{-1} t_l | v_1, ..., v_k \right) \\
&\quad\quad - \text{Cov}\left(t_j \sum v_l^{-1}\ ,\ \sum v_l^{-1} t_l | v_1, ..., v_k \right) + \text{Cov}\left(\sum v_l^{-1} t_l\ ,\ \sum v_l^{-1} t_l | v_1, ..., v_k \right).
\end{aligned}$$

Første ledd er lik null fordi t_i er uavhengig av t_j gitt $v_1, v_2, \ldots, v_k$. Videre er

$$\begin{aligned}
\mathrm{Cov}\left(t_i \sum v_l^{-1}\,,\, \sum v_l^{-1} t_l | v_1, ..., v_k\right) &= \sum v_l^{-1} \mathrm{Cov}\left(t_i\,,\, \sum v_l^{-1} t_l | v_1, ..., v_k\right) \\
&= \sum v_l^{-1} \mathrm{Cov}\left(t_i\,,\, v_i^{-1} t_i | v_1, ..., v_k\right) \\
&\quad + \sum v_l^{-1} \mathrm{Cov}\left(t_i\,,\, \sum_{l \neq i} v_l^{-1} t_l | v_1, ..., v_k\right) \\
&= v_i^{-1} \sum v_l^{-1} \mathrm{Var}\left(t_i | v_1, ..., v_k\right) + 0 \\
&= \sum v_l^{-1}.
\end{aligned}$$

Det følger at også

$$\mathrm{Cov}\left(t_j \sum v_l^{-1}\,,\, \sum v_l^{-1} t_l | v_1, ..., v_k\right) = \sum v_l^{-1}.$$

Det siste leddet kan skrives som

$$\begin{aligned}
\mathrm{Cov}\left(\sum v_l^{-1} t_l\,,\, \sum v_l^{-1} t_l | v_1, ..., v_k\right) &= \mathrm{Var}\left(\sum v_l^{-1} t_l | v_1, ..., v_k\right) \\
&= \sum v_l^{-2} \mathrm{Var}(t_l | v_1, ..., v_k) = \sum v_l^{-1}.
\end{aligned}$$

Innsetting i uttrykket for kovariansen gir at

$$\begin{aligned}
\mathrm{Cov}\left(t_i^*\,,\, t_j^* | v_1, ..., v_k\right) &= \frac{1}{(v_i^*)^{1/2}(v_j^*)^{1/2}(\sum v_l^{-1})^2}\left(0 - 2\sum v_l^{-1} + \sum v_l^{-1}\right) \\
&= -\frac{1}{(v_i^*)^{1/2}(v_j^*)^{1/2}\sum v_l^{-1}}.
\end{aligned}$$

Jeg ønsker å kontrollere at uttrykket er korrekt. Jeg finner $\mathrm{Cov}(t_i^*, t_j^* | v_1, \ldots, v_k)$ ved å nytte en annen framgangsmåte. Vi har at

$$\mathrm{Cov}(t_i^*, t_j^* | v_1, \ldots, v_k) = \mathrm{E}(t_i^* t_j^* | v_1, \ldots, v_k) - \mathrm{E}(t_i^* | v_1, \ldots, v_k)\mathrm{E}(t_j^* | v_1, \ldots, v_k) = \mathrm{E}(t_i^* t_j^* | v_1, \ldots, v_k)$$

fordi $\mathrm{E}(t_i^* | v_1, \ldots, v_k) = 0$. Videre er

$$\begin{aligned}
\mathrm{E}(t_i^* t_j^* | v_1, \ldots, v_k) &= \mathrm{E}\left(\frac{(t_i - \bar{t})}{(v_i^*)^{1/2}} \frac{(t_j - \bar{t})}{(v_j^*)^{1/2}} | v_1, \ldots, v_k\right) \\
&= \frac{1}{(v_i^*)^{1/2}(v_j^*)^{1/2}}\big(\mathrm{E}(t_i t_j | v_1, \ldots, v_k) + \mathrm{E}(\bar{t}^2 | v_1, \ldots, v_k) \\
&\quad - \mathrm{E}(t_i \bar{t} | v_1, \ldots, v_k) - \mathrm{E}(t_j \bar{t} | v_1, \ldots, v_k)\big).
\end{aligned}$$

Fordi t_i er uavhengig av t_j gitt $v_1, v_2, \ldots, v_k$, vil

$$\mathrm{E}(t_i t_j | v_1, \ldots, v_k) = \mathrm{E}(t_i | v_1, \ldots, v_k)\mathrm{E}(t_j | v_1, \ldots, v_k) = \delta^2.$$

Vi vet at

$$\mathrm{E}(t_i\bar{t}|v_1,\ldots,v_k) = \frac{1}{\sum v_l^{-1}} + \delta^2.$$

Fra tidligere utregninger følger det at også

$$\mathrm{E}(\bar{t}^2|v_1,\ldots,v_k) = \frac{1}{\sum v_l^{-1}} + \delta^2.$$

Ved innsetting får vi at

$$\begin{aligned}\mathrm{Cov}(t_i^*\ ,\ t_j^*|v_1,\ldots,v_k) &= \frac{1}{(v_i^*)^{1/2}(v_j^*)^{1/2}}\left(\delta^2 + \frac{1}{\sum v_l^{-1}} + \delta^2 - 2\left(\delta^2 + \frac{1}{\sum v_l^{-1}}\right)\right)\\ &= -\frac{1}{(v_i^*)^{1/2}(v_j^*)^{1/2}\sum v_l^{-1}}.\end{aligned}$$

Dette bekrefter den tidligere beregningen. Kovariansen er ulik null, og t_i^* er ikke uavhengig av t_j^* gitt $v_1, v_2, \ldots, v_k$.

De k parene $(t_1^*, v_1), (t_2^*, v_2), \ldots, (t_k^*, v_k)$ er uavhengige hvis og bare hvis

$$f(t_1^*, v_1, t_2^*, v_2, \ldots, t_k^*, v_k) = f(t_1^*, v_1)f(t_2^*, v_2)\cdots f(t_k^*, v_k).$$

Denne likheten er ikke tilfredsstilt når $\mathrm{Cov}(t_i^*, t_j^*|v_1,\ldots,v_k) \neq 0$. Forutsetningen om uavhengige par brytes.

Effektestimatene standardiseres for å oppnå et sett med estimater som under nullhypotesen kan antas å være iid [5]. Regelen om dobbelforventning gir at

$$\begin{aligned}\mathrm{Cov}(t_i^*\ ,\ t_j^*) &= \mathrm{E}(t_i^*t_j^*) - \mathrm{E}(t_i^*)\mathrm{E}(t_j^*)\\ &= \mathrm{EE}(t_i^*t_j^*|v_1,\ldots,v_k) - \mathrm{EE}(t_i^*|v_1,\ldots,v_k)\mathrm{EE}(t_j^*|v_1,\ldots,v_k)\\ &= \mathrm{EE}(t_i^*t_j^*|v_1,\ldots,v_k) = -\mathrm{E}\left(\frac{1}{(v_i^*)^{1/2}(v_j^*)^{1/2}\sum v_l^{-1}}\right).\end{aligned}$$

Gitt v_i er t_i normalfordelt med varians $v_i > 0$. Dette medfører at $\mathrm{Cov}(t_i^*, t_j^*) \neq 0$. Ut fra situasjonen bør det være klart at $v_1, v_2, \ldots, v_k$ i den siste utregningen skal oppfattes som stokastiske variable, ikke som observerte verdier av stokastiske variable. De standardiserte effektestimatene, t_i^* og t_j^*, er ikke uavhengige stokastiske variable, hverken om variansene er faste konstanter eller stokastiske variable.

Til informasjon er

$$\mathrm{Cor}(t_i^*, t_j^*|v_1, ..., v_k) = \frac{\mathrm{Cov}(t_i^*, t_j^*|v_1,\ldots,v_k)}{\sqrt{\mathrm{Var}(t_i^*|v_1,\ldots,v_k)\mathrm{Var}(t_j^*|v_1,\ldots,v_k)}} = \mathrm{Cov}(t_i^*, t_j^*|v_1,\ldots,v_k).$$

Dette følger siden $\mathrm{Var}(t_i^*|v_1,\ldots,v_k) = \mathrm{Var}(t_j^*|v_1,\ldots,v_k) = 1$.

5.11 Kan $\mathbf{t}^*$ og $\mathbf{v}$ antas uavhengige under nullhypotesen?

Under nullhypotesen skal t_i^* og v_i være uavhengige for $i = 1, 2, \ldots, k$. Jeg har tidligere vist at dette er oppfylt i testsituasjonen til Begg og Mazumdar. Fordi det er et krav om at også de stokastiske parene $(t_1^*, v_1), (t_2^*, v_2), \ldots, (t_k^*, v_k)$ skal være uavhengige, kan vi omskrive nullhypotesen til

$$H_0 : \mathbf{t}^* \text{ og } \mathbf{v} \text{ er uavhengige.}$$

Med $\mathbf{t}^*$ menes hele vektoren som inneholder alle t_i^*, $i = 1, 2, \ldots, k$. Tilsvarende defineres $\mathbf{v}$ som en vektor som inneholder alle v_i for $i = 1, 2, \ldots, k$. Siden $\text{Cov}(t_i^*, t_j^* | v_1, \ldots, v_k)$ ikke er uavhengig av $v_1, v_2, \ldots, v_k$, vil $f(t_i^*, t_j^* | v_1, \ldots, v_k)$ være en funksjon av variansene. Dette er intuitivt, men kan vises formelt i Begg og Mazumdars modellsituasjon ved følgende resonnement.

Fra tidligere vet vi at fordelingen til t_i gitt $v_1, v_2, \ldots, v_k$ er normal med forventning δ og varians v_i. Vi vet også at t_i er uavhengig av t_j, gitt variansene. Det følger at t_i og t_j gitt $v_1, v_2, \ldots, v_k$ er bivariat normalfordelt med forventning

$$\boldsymbol{\mu} = \begin{pmatrix} \delta \\ \delta \end{pmatrix}$$

og kovariansmatrise

$$\boldsymbol{\Sigma} = \begin{pmatrix} v_1 & 0 \\ 0 & v_2 \end{pmatrix}.$$

Dersom variansene er gitte, vil t_i^* og t_j^* være lineære kombinasjoner av uavhengige og normalfordelte t-er. Egenskaper ved den multivariate normalfordelingen gir at også t_i^* og t_j^* har en bivariat normalfordeling. Forventningen er gitt ved

$$\boldsymbol{\mu^*} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Kovariansmatrisen er

$$\boldsymbol{\Sigma}^* = \begin{pmatrix} 1 & -\dfrac{1}{\sum v_l^{-1} (v_i^*)^{1/2} (v_j^*)^{1/2}} \\ -\dfrac{1}{\sum v_l^{-1} (v_i^*)^{1/2} (v_j^*)^{1/2}} & 1 \end{pmatrix}.$$

Simultanfordelingen til t_i^* og t_j^* gitt $v_1, v_2, \ldots, v_k$ er en funksjon av variansene. Ved samme resonnement kan en sette opp hele den simultane betingede fordelingen for $t_1^*, t_2^*, \ldots, t_k^*$ gitt $v_1, v_2, \ldots, v_k$. Denne er definert i Seksjon 6.2. Det følger at $\mathbf{t}^*$ og $\mathbf{v}$ ikke er uavhengige under nullhypotesen.

I Begg og Mazumdars simuleringssituasjon er variansene formelt sett faste størrelser. Da gir det ikke mening å snakke om stokastisk uavhengighet mellom $\mathbf{t}^*$ og $\mathbf{v}$. Vi kan likevel beregne

empiriske korrelasjoner. Matematisk uavhengighet mellom $\mathbf{t}^*$ og $\mathbf{v}$ vil i en slik situasjon være definert ved $f(\mathbf{t}^*|\mathbf{v}) = f(\mathbf{t}^*)$. Definisjonen for matematisk uavhengighet er lik definisjonen for stokastisk uavhengighet, med unntak av at $\mathbf{v}$ ikke er en vektor av stokastiske variable. En kan i slike situasjoner teste nullhypotesen om at $t_1^*, t_2^*, \ldots, t_k^*$ er uavhengige og identisk fordelte ved hjelp av Kendalls tau. Det er da ikke krav om at $(t_1^*, v_1), (t_2^*, v_2), \ldots, (t_k^*, v_k)$ skal ha den samme kontinuerlige, bivariate fordelingen. Jeg vil se nærmere på situasjonen med faste varianser i neste kapittel.

Jeg har lenge trodd at Begg og Mazumdar har vært uoppmerksomme på testmetodens utfordringer. Forfatterne nevner ikke problemene i sin artikkel [8]. I ettertid har jeg erfart at dette ikke er tilfellet. Begg [5] nevner at den empiriske standardiseringen baseres på den estimerte gjennomsnittseffekten, som vil være biased. Dette vil føre til en liten korrelasjon mellom de standardiserte effektestimatene og sampelstørrelsene. Begg utfører ikke beregninger som understøtter dette. Korrelasjon mellom de standardiserte effektestimatene og sampelstørrelsene medfører, i de fleste tilfeller, også en korrelasjon mellom de standardiserte effektestimatene og variansene. Begg [5] hevder videre at dette ikke vil være av særlig betydning fordi testen generelt vil ha lav styrke. Utfordringen med korrelasjon mellom $\mathbf{t}^*$ og $\mathbf{v}$ nevnes kun kort i en parantes, noe som bagatelliserer problemet ytterligere. Jeg vil tilbakevise at konsekvensene er neglisjerbare i Kapittel 7.

Legg merke til at beregningene i den teoretiske delen av nåværende kapittel også vil gjelde om en nytter en rangkorrelasjonstest basert på Spearmans rho. De aktuelle utfordringene er ikke kun knyttet opp mot en rangkorrelasjonstest basert på Kendalls tau.

6 Hvordan påvirkes signifikansnivået til rangkorrelasjonstesten basert på Kendalls tau ved brudd på de ulike forutsetningene?

Ved hjelp av enkle simuleringer vil jeg gi et innblikk i hvordan hver enkelt av de brutte forutsetningene påvirker signifikansnivået til rangkorrelasjonstesten basert på Kendalls tau. Dette vil forhåpentligvis øke forståelsen for problemet med Begg og Mazumdars testmetode, men vil ikke gi et helhetlig bilde av hva som går galt. Simuleringene utdyper ikke hvordan den spesielle korrelasjonen,

$$\mathrm{Cor}(t_i^*, t_j^* | v_1, ..., v_k) = -\frac{1}{(v_i^*)^{1/2}(v_j^*)^{1/2}\sum v_l^{-1}}, \tag{3}$$

virker inn på nivået. Korrelasjonsuttrykket viser at variablene, $t_1^*, t_2^*, \ldots, t_k^*$ og $v_1, v_2, \ldots, v_k$, er korrelerte på kryss og tvers.

Flere forutsetninger brytes når t^* og v korreleres ved hjelp av Kendalls tau. Med samspill eller vekselvirkning menes at en faktor i kombinasjon med en annen påvirker nivået på ulike måter utover det rent additive. Simuleringene utdyper ikke om det er et samspill mellom de brutte forutsetningene, ei heller hvordan en eventuell vekselvirkning virker inn på nivået.

Det er mange ulike simuleringsscenarioer og kombinasjoner av parameterverdier som bør vurderes. Jeg ser kun på noen få. Simuleringsresultatene er ment å veilede, de skal ikke oppfattes som et bevis på noen som helst måte.

6.1 Er nivået ukorrekt fordi Begg og Mazumdar formelt sett lar variansene være faste størrelser?

Begg og Mazumdar [8] lar variansene formelt sett være faste størrelser. Forutsetningen for å utføre en rangkorrelasjonstest brytes. De k parene $(t_1^*, v_1), (t_2^*, v_2), \ldots, (t_k^*, v_k)$ har ikke den samme bivariate fordelingen. Det gir ikke mening å snakke om stokastisk uavhengighet mellom $\mathbf{t}^*$ og $\mathbf{v}$. Dette kan muligens påvirke simuleringsresultatene i Kapittel 4. Det er likevel ikke nødvendig å korrigere forfatternes testmetode grunnet denne problematikken. I virkeligheten er variansene stokastiske variable. Denne forutsetningen vil ikke brytes i reelle situasjoner.

Mange statistiske undersøkelser har form av at samme størrelse observeres på etterfølgende tidspunkter over en viss tid [32]. Kendalls tau, t, kan også brukes til å teste for trend. La X-variablene betegne tid. Måleverdiene på tidspunktene $1, 2, \ldots, n$ kan skrives som $Y_1, Y_2, \ldots, Y_n$ Vi kan da teste nullhypotesen om at $Y_1, Y_2, \ldots, Y_n$ er uavhengige og identisk fordelte ved å beregne t [27]. Her kan vi anse X-variablene, tidspunktene, som faste størrelser. Fordelingen til t under nullhypotesen er nøyaktig den samme som ved rangkorrelasjonstesten basert på

Kendalls tau [27]. Denne testen kalles Mann-Kendalls trendtest. Likheten mellom disse testene gir grunn til å tro at det feilaktige nivået ikke kommer av at Begg og Mazumdar formelt sett lar variansene være faste størrelser.

Jeg konstruerer enkle simuleringseksempler for å understøtte denne antakelsen. Jeg velger $k = 75$ og lar variansene være faste størrelser, $v = 0.1, 1.0, 10.0$. For hver varians tillegges $\epsilon = 0.0001$ for å unngå ties. Deretter genereres $k = 75$ uavhengige standardiserte effektestimater, $t_1^*, t_2^*, \ldots, t_{75}^*$, fra en standardnormalfordeling. De er altså "standardiserte" på forhånd slik at $f(\mathbf{t}^*|v_1, \ldots, v_k)$ ikke er en funksjon av variansene. Kendalls tau beregnes ved å korrelere t^* og v, og simuleringsprosedyren gjentas 10000 ganger. Jeg beregner det empiriske signifikansnivået for rangkorrelasjonstesten ved et tosidig nominelt nivå på 0.05. Nivået estimeres til 0.0490. Tosidige konfidensintervaller med konfidensnivå 0.95 gir ikke grunnlag til å forkaste hypotesen om at signifikansnivået er lik 0.05.

Prosedyren ovenfor gjentas. Den eneste forskjellen er at de standardiserte effektestimatene genereres fra en kontinuerlig uniform fordeling på intervallet [0,1]. Variansene er som før. Det empiriske nivået er 0.0505. Jeg gjentar også simuleringene når t_i^* genereres fra en Gamma(0.1,1)-fordeling, $i = 1, 2, \ldots, 75$. Nivået estimeres til 0.0519. Vi forkaster heller ikke for disse situasjonene hypotesen om at signifikansnivået er lik det nominelle, basert på tosidige konfidensintervaller med konfidensnivå 0.95. Simuleringene underbygger at faste varianser ikke er skyld i det dårlige nivået.

Vil nivået til Begg og Mazumdars testmetode være korrekt i situasjoner hvor variansene er stokastiske variable? Simuleringer avkrefter dette. Jeg gjentar simuleringsprosedyren gitt av Begg og Mazumdar som beskrevet i Kapittel 3, under nullhypotesen. Den underliggende effektparameteren settes lik null. Variansene trekkes derimot fra en invers gammafordeling med formparameter $\alpha = 100$ og skalaparameter $\beta = 1$. Jeg lar $k = 75$. Effektestimatene genereres gitt variansene og standardiseres. Kendalls tau beregnes ved å korrelere de standardiserte effektestimatene og variansene, slik at korrelasjonen gitt ved Likning (3) innføres. Simuleringene gjentas 10000 ganger. Nye varianser trekkes fra den inverse gammafordelingen for hver repetisjon. Det tosidige nominelle signifikansnivået settes også denne gangen lik 0.05. Nivået estimeres til 0.0535. Vi forkaster ikke hypotesen om at signifikansnivået er lik det nominelle ved et tosidig 5%-nivå. Simuleringene gjentas også med varianser fra en invers gammafordeling, hvor $\alpha = 0.1$ og $\beta = 1$. Her er variansspredningen større. Nivået estimeres til 0.0131. Dette er for lavt. Faste varianser er uansett ikke den eneste årsaken til det feilaktige nivået.

Den empiriske standardiseringen baseres på den estimerte gjennomsnittseffekten. Begg [5] hevder dette er opphavet til korrelasjonen som oppstår mellom de standardiserte effektestimatene og variansene. I simuleringsscenarioene beskrevet av Begg og Mazumdar [8] vet vi den sanne effekten. Jeg baserer standardiseringen på denne underliggende effekten og repeterer

simuleringene beskrevet i Kapittel 3 under nullhypotesen. Simuleringsresultatene bekrefter at signifikansnivået er tilnærmet lik det nominelle. Disse resultatene refereres ikke i oppgaven. Simuleringsresultatene understøtter at årsaken til det lave nivået er knyttet til standardiseringen av effektestimatene og korrelasjonen, gitt ved Likning (3), som denne standardiseringen medfører. Denne korrelasjonen forårsaker brudd på to av forutsetningene for å utføre en rangkorrelasjonstest basert på Kendalls tau. Jeg vil videre undersøke konsekvenser av brudd på hver enkelt av disse forutsetningene. Dette er selvsagt noe forenklet. Begg og Mazumdars testmodell innfører, som tidligere bemerket, korrelasjon på kryss og tvers av de stokastiske variablene som inngår i rangkorrelasjonstesten. Simuleringene vil likevel gi nyttig informasjon om rangkorrelasjonstesten basert på Kendalls tau.

6.2 Hvordan påvirkes nivået av at de forskjellige observasjonsparene ikke er uavhengige?

Den simultane fordelingen til $t_1^*, t_2^*, \ldots, t_k^*$ gitt $v_1, v_2, \ldots, v_k$ er multivariat normalfordelt i testsituasjonen til Begg og Mazumdar. Forventningen er gitt ved $\underset{k\times 1}{\boldsymbol{\mu}}{}^* = (0, 0, \ldots, 0)^T$, hvor symbolet T betegner transponering av matrisen. Korrelasjonsmatrisen er

$$\underset{k\times k}{\boldsymbol{\Sigma}}{}^* = \begin{pmatrix} 1 & -\frac{1}{\sum v_l^{-1}(v_1^*)^{1/2}(v_2^*)^{1/2}} & \cdots & -\frac{1}{\sum v_l^{-1}(v_1^*)^{1/2}(v_k^*)^{1/2}} \\ -\frac{1}{\sum v_l^{-1}(v_2^*)^{1/2}(v_1^*)^{1/2}} & 1 & \cdots & -\frac{1}{\sum v_l^{-1}(v_2^*)^{1/2}(v_k^*)^{1/2}} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{\sum v_l^{-1}(v_k^*)^{1/2}(v_1^*)^{1/2}} & -\frac{1}{\sum v_l^{-1}(v_k^*)^{1/2}(v_2^*)^{1/2}} & \cdots & 1 \end{pmatrix}.$$

For alle simuleringer i denne seksjonen beregnes det empiriske signifikansnivået for rangkorrelasjonstesten ved et tosidig nominelt nivå på 0.05. Antall repetisjoner settes lik 10000. I denne seksjonen lar jeg videre $\mathbf{t}^*$ og $\mathbf{v}$ være uavhengige. Jeg ønsker derimot en korrelasjon mellom t_i^* og t_j^*, $i \neq j$.

Det er nærliggende å generere $\mathbf{t}^*$ fra en multivariat normalfordeling med forventningsvektoren og korrelasjonsmatrisen definert ovenfor. Dette gjøres i R ved hjelp av statistikkpakken MASS [53]. Jo større sampelstørrelse, desto sikrere kan en være at den beregnede verdien til observatoren gir et nøyaktig estimat for den underliggende parameterverdien [48]. Vi har i Seksjon 5.1 sett at nivået blir for lavt når Kendalls tau beregnes for 25 observasjonspar. Av den grunn velger jeg $k = 75$. Korrelasjonsmatrisen beregnes gitt $v = 0.1, 1.0, 10.0$, og $\mathbf{t}^*$ genereres. For å unngå at $f(\mathbf{t}^*|\mathbf{v})$ er en funksjon av $\mathbf{v}$, trekker jeg *nye*, uavhengige varianser fra en kontinuerlig fordeling. Deretter korreleres de standardiserte effektestimatene og de *nye* variansene ved

hjelp av Kendalls tau. Jeg presiserer at jeg ikke korrelerer de standardiserte effektestimatene mot variansene som brukes til å beregne korrelasjonsmatrisen. De *nye* variansene trekkes først fra en standardnormalfordeling. Effektestimatenes varianser kan i virkeligheten selvsagt ikke være negative. Rangkorrelasjonstesten er skaleringsinvariant, og negative varianser vil ikke påvirke resultatene i denne testsammenhengen. Nye varianser trekkes for hver av de 10000 repetisjonene. Nivået estimeres til 0.0518. Jeg trekker deretter *nye* varianser fra en gammafordeling med parametre $\alpha = 1$ og $\beta = 1$, men gjentar ellers den overnevnte prosessen. Det estimerte nivået er 0.0497. Jeg forkaster ikke nullhypotesen om at signifikansnivået er lik det nominelle for noen av tilfellene ved et tosidig nivå på 5%. Korrelasjonen som innføres mellom t_i^* og t_j^* ser ikke ut til å påvirke nivået i vesentlig grad.

Noen vil kanskje innvende at dette ikke er overraskende. For alle $i \neq j$, vil $-0.03779 < \text{Cor}(t_i^*, t_j^* | v_1, \ldots, v_k) < 0$ dersom $v = 0.1, 1.0, 10.0$ og $k = 75$. Det er likevel nærliggende å undersøke om korrelasjon mellom variablene kan ødelegge nivået. Dersom $v = 0.1, 1.0, 10.0$, men $k = 25$, vil $-0.12569 < \text{Cor}(t_i^*, t_j^* | v_1, \ldots, v_k) < 0$. De største korrelasjonsverdiene er ikke neglisjerbare.

Jeg gjentar simuleringene ovenfor, men lar $\text{Cor}(t_i^*, t_j^*) = 0.8$ for alle $i \neq j$, $i, j = 1, 2, \ldots, 75$. De standardiserte effektestimatene genereres fra en multivariat normalfordeling med denne korrelasjonsmatrisen. Forventningsvektoren er gitt innledningsvis i denne seksjonen. Nivået estimeres til 0.0483 når de standardiserte effektestimatene korreleres mot uavhengige standardnormalfordelte varianser. Jeg gjentar simuleringsprosedyren, men korrelerer de standardiserte effektestimatene mot varianser trukket fra en Gamma$(1, 1)$-fordeling. Det empiriske nivået er 0.049.

Nå trekker jeg 75 varianser fra en Gamma$(1, 1)$-fordeling. Gitt disse verdiene, beregner jeg korrelasjonsmatrisen og genererer $t_1^*, t_2^*, \ldots, t_{75}^*$. De standardiserte effektestimatene korreleres mot uavhengige standardnormalfordelte varianser, slik at $\mathbf{t}^*$ og $\mathbf{v}$ er uavhengige. Nivået estimeres til 0.0497.

Hva om også $v_1, v_2, \ldots, v_k$ er korrelerte? Korrelasjonsmatrisen inngår eksplisitt i den multivariate normalfordelingen. Det er lett å generere variansene fra en multivariat normalfordeling med ønsket korrelasjonsmatrise, selv om variansene i virkeligheten ikke har en slik fordeling. Jeg setter $v = 0.1, 1.0, 10.0$ og beregner korrelasjonsmatrisen gitt innledningsvis i denne seksjonen. Både de standardiserte effektestimatene og variansene genereres fra en multivariat normalfordeling med denne korrelasjonsmatrisen og forventningsvektoren gitt innledningsvis. Deretter beregnes Kendalls tau ved å korrelere t^* og v. Igjen velges $k = 75$. Det empiriske nivået er 0.0511. Jeg gjentar simuleringene. Nå er korrelasjonsmatrisen til de standardiserte effektestimatene gitt ved $\text{Cor}(t_i^*, t_j^*) = 0.8$, $i, j = 1, 2, \ldots, 75, i \neq j$. Korrelasjonsmatrisen til variansene er gitt ved $\text{Cor}(v_i, v_j) = 0.8$, $i, j = 1, 2, \ldots, 75, i \neq j$. Nivåestimatet er 0.0488.

Heller ikke for disse situasjonene forkastes nullhypotesen om at signifikansnivået er lik det nominelle, basert på tosidige konfidensintervaller med konfidensnivå 0.95.

Jeg ønsker å presisere at jeg på ingen måte har vist at korrelerte observasjonspar ikke påvirker nivået i en rangkorrelasjonstest. Jeg har i hovedsak tatt sikte på å undersøke den praktiske betydningen av ikke-uavhengige observasjonspar i eksempler som kan minne om situasjonen vi har i Begg og Mazumdars artikkel [8]. I simuleringene har jeg derfor antatt at simultanfordelingen til $t_1^*, t_2^*, \ldots, t_k^*$ er multivariat normal.

Jeg kan ikke utelukke at korrelasjon mellom observasjonsparene vil påvirke det estimerte nivået dersom $t_1^*, t_2^*, \ldots, t_k^*$ har en annen simultanfordeling. En kan bruke copulas til å generere $t_1^*, t_2^*, \ldots, t_k^*$ med en annen simultanfordeling enn den multivariat normale. Eksempelvis kan alle t_i^* ha den samme marginale fordelingen, samtidig som t_i^* er korrelert med t_j^*, $i \neq j$. Resultater fra slike simuleringer kan gi interessant informasjon om testen basert på Kendalls tau. De vil likevel ikke gi relevant informasjon i forhold til å finne problemet med Begg og Mazumdars simuleringsresultater [8]. Slike simuleringer gjennomføres derfor ikke i denne oppgaven.

Jeg vil ikke forsøke å modifisere testprosedyren til Begg og Mazumdar [8] i tilfeller hvor en har uavhengighet mellom $\mathbf{t}^*$ og $\mathbf{v}$, men ikke uavhengige observasjonspar. En slik modifikasjon vil, i situasjoner hvor effektestimatene antas normalfordelte, mest sannsynlig kun være av teoretisk betydning.

6.3 Hvordan påvirkes nivået dersom t_i^* og v_i ikke er uavhengige under nullhypotesen?

Jeg genererer 75 varianser fra en Gamma(1,1)-fordeling. Ut fra disse verdiene, beregnes korrelasjonsmatrisen gitt i forrige seksjon. De standardiserte effektestimatene, $t_1^*, t_2^*, \ldots, t_{75}^*$, genereres fra en multivariat normalfordeling med denne korrelasjonsmatrisen. Forventningen er fortsatt nullvektoren. De standardiserte effektestimatene korreleres ved hjelp av Kendalls tau mot de *samme* variansene som ble brukt til å regne ut korrelasjonsmatrisen. Dette gir en liknende korrelasjon mellom $\mathbf{t}^*$ og $\mathbf{v}$ som den vi får ved utføring av Begg og Mazumdars testmetode. Prosedyren gjentas 10000 ganger. Det empiriske nivået for rangkorrelasjonstesten beregnes ved et tosidig nominelt nivå på 0.05. Et tosidig nominelt nivå på 0.05 velges også for de andre simuleringsscenarioene i denne seksjonen. Nivået estimeres til 0.0248. Jeg forkaster hypotesen om at signifikansnivået er lik 0.05, basert på et tosidig konfidensintervall med konfidensnivå 0.95. Korrelasjon mellom $\mathbf{t}^*$ og $\mathbf{v}$ virker inn på nivået til Begg og Mazumdars testmetode. Hvordan er fortsatt uklart, særlig fordi jeg ikke kan utelukke vekselvirkning.

De stokastiske variablene t_i^* og v_i, $i = 1, 2, \ldots, k$, vil være uavhengige i Begg og Mazumdars testsituasjon. Med tanke på den generelle rangkorrelasjonstesten basert på Kendalls tau, gir det likevel mening å undersøke hvordan korrelasjon mellom t_i^* og v_i påvirker nivået. Jeg innfører

ulike simuleringssituasjoner hvor t_i^* ikke er uavhengig av v_i.

La de 75 parene $(t_1^*, v_1), (t_2^*, v_2), \ldots, (t_{75}^*, v_{75})$ være uavhengige. Parene er alle bivariat normalfordelte. Forventningen og korrelasjonsmatrisen innad i hvert par er gitt ved henholdsvis $\boldsymbol{\mu} = (0,0)^T$ og

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}.$$

Jeg korrelerer t^* og v ved hjelp av Kendalls tau. Her er alle forutsetninger oppfylt, bortsett fra at t_i^* ikke er uavhengig av v_i under nullhypotesen, $i = 1, 2, \ldots, k$. Prosedyren gjentas 10000 ganger. Nivåestimatet er 0.1268. Brudd på forutsetningen om uavhengighet under nullhypotesen kan få alvorlige konsekvenser.

Jeg vurderer også andre avhengighetssituasjoner. La $\mathbf{v}$ bestå av alle heltall på intervallet [-25,25]. Jeg innfører monoton sammenheng mellom t_i^* og v_i ved formelen $t_i^* = v_i + \mathrm{N}(0, \sigma^2)$. Jeg korrelerer t^* og v ved hjelp av Kendalls tau. Prosessen gjentas 10000 ganger. Styrken estimeres. Dersom $\sigma = 0$, er følgelig styrken 1. Dersom $\sigma = 100$, estimeres den til 0.1599. Videre gir $\sigma = 1000$ og $\sigma = 10000$ styrkeestimater på henholdsvis 0.0516 og 0.0517. Styrken stabiliserer seg rundt 0.05 når σ blir stor. For liten σ vil t_i^* og v_i være sterkt empirisk korrelerte. Denne korrelasjonen avtar desto større σ blir.

Prosessen i det forrige avsnittet gjentas. Jeg definerer nå $t_i^* = v_i^2 + \mathrm{N}(0, \sigma^2)$, mens variansene er som før. Sammenhengen mellom t_i^* og v_i er verken lineær eller monoton. Styrken er lik 0 dersom $\sigma = 0$. Den er tilnærmet lik null for $\sigma = 100$. For $\sigma = 1000$, estimeres styrken til 0.0434. Her forkastes hypotesen om at nivået er lik 0.05 ved et tosidig 5%-nivå. Styrken stabiliserer seg rundt 0.05 for store verdier av σ og estimeres til 0.0498 når $\sigma = 10000$.

Simuleringsresultatene bekrefter at signifikansnivået kan være både lavere eller høyere enn det nominelle dersom $f(\mathbf{t}^* | v_1, \ldots, v_k)$ er en funksjon av $v_1, v_2, \ldots, v_k$ under nullhypotesen. Det er betryggende at rangkorrelasjonstesten basert på Kendalls tau gir utslag når jeg innfører en monoton sammenheng mellom t_i^* og v_i. Hvis ikke ville testen vært ubrukelig til sitt formål.

Sheskin [48] definerer nullhypotesen til Kendalls tau ved

$$H_0 : \tau = 0.$$

Som tidligere bemerket, misliker jeg denne notasjonen. Simuleringsresultatene lar meg utdype hvorfor. Selv om sammenhengen mellom t_i^* og v_i verken er lineær eller monoton under nullhypotesen for $i = 1, 2, \ldots, k$, kan signifikansnivået være ulik det nominelle dersom t_i^* og v_i ikke er uavhengige. Til tross for at $\tau = 0$ under nullhypotesen, risikerer vi et feilaktig nivå. Den asymptotiske fordelingen til Kendalls tau under nullhypotesen er utledet ved uavhengige stokastiske variable. Bevisene vil ikke holde om vi letter på kravene og kun krever at de stokastiske variablene skal være ukorrelerte.

7 Forslag til mulig forbedring av Begg og Mazumdars testmetode

I dette kapittelet vil jeg foreslå en mulig forbedring av testprosedyren til Begg og Mazumdar [8]. Også denne testmetoden baseres på rangkorrelasjoner. Først vil jeg gi en kort, generell innføring om hypotesetesting og teori rundt sammenlikning av ulike testmetoder. Noe av denne teorien danner motivasjonen for testprosedyren jeg foreslår. Deretter vil jeg vurdere testmodellen min opp mot den introdusert av Begg og Mazumdar.

7.1 Kort om hypotesetesting og evaluering av ulike testmetoder

En hypotesetest er en observasjonsbasert metode for å klargjøre om en gitt hypotese vedrørende en statistisk modell, kalt nullhypotesen, bør forkastes til fordel for en gitt alternativ hypotese [32]. Nullhypotesen betegnes H_0, mens H_1 betegner den alternative hypotesen. Vi ønsker altså å teste nullhypotesen, $H_0 : \theta \in \Theta_0$, mot den alternative hypotesen, $H_1 : \theta \in \Theta_0^c$, hvor θ er en parameter. Videre er Θ_0 en undermengde av parameterrommet med komplement Θ_0^c. En testmetode defineres ved en testobservator og et forkastningsområde [32]. En testobservator er en stokastisk variabel, knyttet til målingene, som vi kan basere vår beslutning på. Det er viktig at sannsynlighetsfordelingen til testobservatoren beskrives av parameteren θ [34]. Forkastningsområdet er de verdiene av testobservatoren som medfører at nullhypotesen skal forkastes.

En kan ofte velge mellom flere testmetoder når en skal teste en hypotese. Når en bestemmer seg for å forkaste eller ikke å forkaste nullhypotesen, kan det hende en trekker gal slutning. Hypotesetester kan evalueres og sammenliknes ved å vurdere testenes sannsynlighet for å gjøre feil.

En hypotesetest har to feiltyper. En feil av type I forekommer dersom $\theta \in \Theta_0$, og hypotesetesten feilaktig forkaster H_0. Det kan også hende at $\theta \in \Theta_0^c$, samtidig som testmetoden velger ikke å forkaste H_0. Dette defineres som en feil av type II.

La $X_1, X_2, \ldots, X_n$ være uavhengige og identisk fordelte stokastiske variable. Vi observerer $x_1, x_2, \ldots, x_n$. La R være hypotesetestens forkastningsområde. Hvis $\theta \in \Theta_0$, vil testen gjøre feil dersom $\mathbf{x} \in R$. Sannsynligheten for en feil av type I er $P_\theta(\mathbf{X} \in R)$. Dersom $\theta \in \Theta_0^c$, er sannsynligheten for en feil av type II $P_\theta(\mathbf{X} \in R^c) = 1 - P_\theta(\mathbf{X} \in R)$ [10]. Tabell 10 oppsummerer definisjonene.

Videre er det nødvendig å definere hva som menes med en styrkefunksjon. Styrkefunksjonen til en hypotesetest med forkastningsområde R er en funksjon av θ, definert ved $\beta(\theta) = P_\theta(\mathbf{X} \in R)$ [10].

Det ideelle er selvsagt ikke å gjøre feil. Dette tilsvarer en styrkefunksjon som er 0 for alle

Tabell 10: Feiltyper ved hypotesetesting

		Beslutning	
		Ikke forkast H_0	Forkast H_0
Sann hypotese	H_0	Korrekt beslutning	Feil av type I
	H_1	Feil av type II	Korrekt beslutning

$\theta \in \Theta_0$ og 1 for alle $\theta \in \Theta_0^c$. En slik styrkefunksjon oppnår man kun i trivielle situasjoner. En god test har en styrkefunksjon som kommer nær 0 for de fleste $\theta \in \Theta_0$ og nær 1 for de fleste $\theta \in \Theta_0^c$ [10]. Testen har da både lav feil av type I og lav feil av type II.

Verdien til en styrkefunksjon for en gitt $\theta \in \Theta_0^c$ kalles for styrken til testen i dette punktet. Dersom

$$\max_{\theta \in \Theta_0} \beta(\theta) = \alpha,$$

sier vi at testen har signifikansnivå α [32]. Jeg vil straks presentere en noe mer komplisert definisjon av en test med signifikansnivå α, som passer bedre til den videre diskusjonen i oppgaven.

Styrkefunksjonen til en test vil typisk avhenge av sampelstørrelsen n [10]. Generelt vil et lavere signifikansnivå medføre større feil av type II dersom sampelstørrelsen holdes konstant og vi holder oss til den samme testobservatoren. Motsatt vil lavere feil av type II gi større feil av type I [22]. Ved å øke n kan en imidlertid oppnå lavere feil av både type I og type II. Dersom en ønsker å velge mellom flere tester, er den vanlige prosedyren å begrense seg til tester som kontrollerer feil av type I på et gitt nivå. Innenfor denne begrensede klassen av tester, leter man etter den testen som har lavest feil av type II [10].

Ved sammenlikning av tester, er det viktig på en eller annen måte å kunne kontrollere feil av type I. Det har lite for seg å lete etter den testen som inneholder lavest feil av type II dersom det ikke finnes restriksjoner på feil av type I. Eksempelvis vil en test som forkaster H_0 med sannsynlighet 1 aldri kunne gjøre feil av type II.

For en videre diskusjon trenger jeg flere definisjoner. En test med styrkefunksjon $\beta(\theta)$ er en test med størrelse α dersom $\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha$ for $0 \leq \alpha \leq 1$. En test med styrkefunksjon $\beta(\theta)$ er en test med nivå α dersom $\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$, $0 \leq \alpha \leq 1$ [10]. Mengden av nivå α-tester inneholder mengden av størrelse α-tester.

En foretrekker tester med størrelse α. Beregningsmessig kan det ofte være vanskelig å konstruere slike tester. Da ser en seg ofte fornøyd med nivå α-tester, selv om dette gir tester som er for konservative [10].

Ved hypotesetesting er vanlig framgangsmåte å spesifisere testens signifikansnivå. Da kontrollerer man feil av type I, ikke feil av type II. Nullhypotesen og den alternative hypotesen

bør være definert på en slik måte at det er viktigere å kontrollere feil av type I enn feil av type II. Casella og Berger [10] gir et forklarende eksempel. En forventer at et eksperiment skal støtte opp om en gitt hypotese. Derimot ønsker en ikke å komme med denne påstanden dersom dataene ikke gir overbevisende støtte. Den alternative hypotesen bør være den gitte hypotesen en forventer at dataene skal støtte opp om, og som en håper å bevise. Ved å bruke en nivå α-test med liten α, garderer en seg mot å si at dataene støtter opp om den alternative hypotesen dersom denne er falsk.

Hensikten med statistisk hypotesetesting er å ta stilling til om det er overveiende sannsynlig at den alternative hypotesen, H_1, er riktig [34].

Når en vurderer tester, bør en også ta andre hensyn i tillegg til å kontrollere feil av type I på et gitt nivå. Vi ønsker en test som forkaster H_0 med større sannsynlighet dersom $\theta \in \Theta_0^c$ enn hvis $\theta \in \Theta_0$. En test med denne egenskapen er styrkerett. Formelt er en styrkefunksjon $\beta(\theta)$ styrkerett dersom $\beta(\theta') \geq \beta(\theta'')$ for alle $\theta' \in \Theta_0^c$ og $\theta'' \in \Theta_0$ [10].

En kan også nytte beslutningsteori for å vurdere og sammenlikne ulike testmetoder. Da kan en ta hensyn til om en feil er mer alvorlig enn en annen. Risikofunksjonen, definert som forventet feil, kan brukes til å evaluere tester. Jeg gir en kort innføring i beslutningsteori. Samtlige av de resterende definisjonene i denne seksjonen er hentet fra Casella og Berger [10].

Ved hypotesetesting kan vi utføre to mulige handlinger, ikke forkaste nullhypotesen eller forkaste nullhypotesen. Jeg betegner disse beslutningene henholdsvis a_0 og a_1. Handlingsrommet ved hypotesetesting er mengden $\mathcal{A} = \{a_0, a_1\}$. Vi kan beskrive en beslutningsregel $\delta(\mathbf{x})$, også kalt en hypotesetest, som en funksjon på $\mathcal{X}$ som kun kan ta de to verdiene a_0 og a_1. Her betegner $\mathcal{X}$ utfallsrommet til $\mathbf{X}$. Mengden $\{\mathbf{x} : \delta(\mathbf{x}) = a_1\}$ er testens forkastningsområde. Mengden $\{\mathbf{x} : \delta(\mathbf{x}) = a_0\}$ beskriver området hvor en ikke forkaster nullhypotesen.

En tapsfunksjon defineres. Tapsfunksjonen bør reflektere at en feil blir gjort hvis $\theta \in \Theta_0$ og nullhypotesen forkastes, eller dersom $\theta \in \Theta_0^c$ og nullhypotesen ikke forkastes. Tapsfunksjonen bør også reflektere at den korrekte beslutningen tas i de to resterende tilfellene.

Tapsfunksjonen for ulike verdier av θ dersom en ikke forkaster nullhypotesen defineres ved $L(\theta, a_0)$. Tapsfunksjonen for ulike verdier av θ dersom en forkaster nullhypotesen er gitt ved $L(\theta, a_1)$. Generalisert 0-1-tap er en realistisk tapsfunksjon som gir ulike feil forskjellige kostnader. Denne defineres ved

$$L(\theta, a_0) = \begin{cases} 0 & \theta \in \Theta_0 \\ c_{\mathrm{II}} & \theta \in \Theta_0^c \end{cases}$$

og

$$L(\theta, a_1) = \begin{cases} c_{\mathrm{I}} & \theta \in \Theta_0 \\ 0 & \theta \in \Theta_0^c \end{cases}.$$

Kostnaden til en feil av type I er c_{I}, mens c_{II} er kostnaden til en feil av type II.

La $\beta(\theta)$ være styrkefunksjonen til testen basert på beslutningsregelen δ. Forkastningsområdet til testen betegnes ved $R = \{\mathbf{x} : \delta(\mathbf{x}) = a_1\}$. Da har vi at $\beta(\theta) = P_\theta(\mathbf{X} \in R) = P_\theta\big(\delta(\mathbf{X}) = a_1\big)$. Risikofunksjonen, $R(\theta, \delta)$, assosiert med denne tapsfunksjonen defineres ved

$$R(\theta, \delta) = 0 \cdot P_\theta\big(\delta(\mathbf{X}) = a_0\big) + c_{\mathrm{I}} P_\theta\big(\delta(\mathbf{X}) = a_1\big) = c_{\mathrm{I}}\beta(\theta) \quad \text{hvis } \theta \in \Theta_0,$$

og

$$R(\theta, \delta) = c_{\mathrm{II}} P_\theta\big(\delta(\mathbf{X}) = a_0\big) + 0 \cdot P_\theta\big(\delta(\mathbf{X}) = a_1\big) = c_{\mathrm{II}}\big(1 - \beta(\theta)\big) \quad \text{hvis } \theta \in \Theta_0^c.$$

Styrkefunksjonen spiller en viktig rolle også her.

Andre valg av tapsfunksjoner er selvsagt mulig. Generalisert 0-1-tap vil ofte være for enkel i mange reelle situasjoner. Det finnes også tapsfunksjoner som ikke kun tar for seg om en avgjørelse er rett eller gal, men som kan vekte hvor alvorlig denne feilen er. En bør velge en tapsfunksjon som passer overens med formålet.

Dersom $\delta_1(\mathbf{x})$ er en hypotesetest og $\delta_2(\mathbf{x})$ en annen, velger vi δ_1 dersom $R(\theta, \delta_1) < R(\theta, \delta_2)$ for alle $\theta \in \Theta$. Ofte vil disse funksjonene krysse hverandre. Da blir valget vanskeligere. Også her kan det være naturlig å kontrollere feil av type I på et gitt nivå og deretter sammenlikne risikofunksjonen for verdier av θ som ligger i den alternative hypotesen. Kontroll av feil av type I tilsvarer kontroll av risikofunksjonen for $\theta \in \Theta_0$.

7.2 Motivasjon og forslag til mulig forbedring av testmetode

På bakgrunn av teori om hypotesetesting er det nok mange som ikke reagerer nevneverdig over at signifikansnivået er lavere enn det nominelle i simuleringsscenarioene presentert av Begg og Mazumdar. I kompliserte testsituasjoner er det ofte beregningsmessig umulig å konstruere størrelse α-tester. En ser seg fornøyd med en test med nivå α [10]. Konservative tester foretrekkes. I denne seksjonen vil jeg derimot argumentere for at en test med nivå α ikke vil være konservativ i en metaanalysesammenheng. Hovedårsaken ligger i hvordan nullhypotesen og den alternative hypotesen er definert.

Nullhypotesen, $H_0 : \mathbf{t}^*$ og $\mathbf{v}$ er uavhengige, er forenelig med nullhypotesen definert ved

$$H_0 : \text{ Metaanalysen inneholder ikke publikasjonsbias.}$$

På samme måte kan den alternative hypotesen, $H_1 : \tau \neq 0$, uttrykkes ved

$$H_1 : \text{ Metaanalysen inneholder publikasjonsbias.}$$

Vi ønsker en testmetode som innebærer at risikoen for begge typer feilslutninger er liten. Disse ønskene trekker derimot i hver sin retning. I praksis må en foreta en rimelig avveining mellom de to ønskene. Alternativt kan man finne ut hvor omfattende undersøkelsen må være

for at begge risikoer skal være redusert til et akseptabelt nivå [32]. Metaanalyser skal inkludere alle aktuelle studier som viser til resultater om en gitt problemstilling. En kan ikke lett øke sampelstørrelsen og må derfor vurdere de ulike feiltypene.

Vi ønsker selvsagt metaanalyser som ikke inneholder publikasjonsbias. Et lavere nivå medfører generelt også lavere styrke. Dersom man forkaster nullhypotesen for sjelden, risikerer en derfor å gjennomføre for mange metaanalyser hvor en ikke kan utelukke publikasjonsbias. Dette kan gi alvorlige konsekvenser og bør unngås. Resultater fra metaanalysen kan være biased. På bakgrunn av dette risikerer en å trekke feilaktige konklusjoner.

Hva om en forkaster nullhypotesen for ofte? Mistanke om publikasjonsbias kan medføre at for mange metaanalyser uten publikasjonsbias ikke utføres. Til gjengjeld minker sannsynligheten for å utføre metaanalyser som er utsatt for selektiv publikasjon. Ved forkastning kan en eventuelt nytte metoder for fellesestimering som forsøker å korrigere for selektiv publikasjon. Trim and Fill [13, 14] er en iterativ og ikke-parametrisk metode som gir et korrigert estimat for behandlingseffekten. Duval og Tweedie antar at seleksjon av studier avhenger av behandlingseffektens størrelse og retning. Metoden er basert på symmetriegenskapene til funnelplottet. Etter mitt skjønn er feil av type I mindre alvorlig enn å utføre metaanalyser hvor resultatene kan være biased grunnet publikasjonsbias. Metoder for fellesestimering som forsøker å korrigere for selektiv publikasjon er et sentralt tema innen publikasjonsbias i metaanalyser. Dette er er tema for videre arbeid. Jeg vil dessverre ikke arbeide med slike metoder i denne oppgaven.

Det er altså viktigere å kontrollere feil av type II enn feil av type I. Dette er forskjellig fra de fleste situasjoner jeg møter i statistiske sammenhenger. I henhold til teorien beskrevet i Seksjon 7.1, bør hypotesene redefineres. Den alternative hypotesen velges som den arbeidshypotesen en ønsker å teste og den påstanden som krever bevis. Nullhypotesen velges som den motsatte påstanden. Vi forventer at dataene støtter opp om hypotesen om ingen publikasjonsbias. En passende formulering av nullhypotesen vil være

$$H_0: \text{ Metaanalysen inneholder publikasjonsbias.}$$

Den alternative hypotesen bør defineres ved

$$H_1: \text{ Metaanalysen inneholder ikke publikasjonsbias.}$$

Styrkefunksjonen er kontinuerlig. Det er ikke mulig å konstruere tester med rimelig styrke for disse definisjonene. Muligens burde jeg definert nullhypotesen ved

$$H_0: \text{ Ikke-neglisjerbar bias}$$

og den alternative hypotesen ved

$$H_1: \text{Neglisjerbar bias.}$$

Betydningen av ordet "neglisjerbar" kan selvsagt diskuteres. For disse hypotesedefinisjonene kan en mest sannsynlig oppnå god styrke for store datasett. I metaanalyser kan en ikke uten videre øke sampelstørrelsen. En redefinering av hypotesene vil ikke være en god idé ut fra den praktiske sammenhengen.

Jeg holder meg til hypotesene slik de opprinnelig er definert. Jeg forsøker å forbedre Begg og Mazumdars testmetode ut fra disse definisjonene. Teorien om hypotesetesting må brukes med forsiktighet. Hypotesene er ikke definert slik at feil av type I foretrekkes framfor feil av type II.

Begg [5] er også klar over at konsekvensene av en falsk negativ test er mer alvorlige enn en falsk positiv test i en metaanalysesammenheng. Ifølge Begg bør en velge et liberalt signifikansnivå når en tester for publikasjonsbias. Dette er jeg selvsagt enig i, men hverken de teoretiske eller praktiske problemene knyttet til testprosedyren introdusert av Begg og Mazumdar vil forsvinne om signifikansnivået liberaliseres. Dette bekreftes i Kapittel 11. Her estimeres signifikansnivået ved et nominelt nivå på 0.10 i en annerledes simuleringssituasjon. Flere tiltak må til.

Begg og Mazumdar [8] utfører tilleggssimuleringer hvor de korrelerer effektestimatene og variansene under nullhypotesen ved et tosidig nominelt nivå på 0.05. Effektestimatene standardiseres ikke her, i motsetning til tidligere situasjoner. De rapporterer en gjennomsnittlig teststørrelse på 0.10 for $k = 25$ og stor variansspredning. For $k = 25$ og liten variansspredning er den gjennomsnittlige teststørrelsen 0.09. Ifølge Begg og Mazumdar demonstrerer resultatene at testmetoden uten standardiserte effektestimater ikke er gyldig. Jeg er uenig. I tradisjonell statistikk virker man særlig opptatt av å begrense sannsynligheten for feil av type I ved et gitt nivå. Når en tester for publikasjonsbias i metaanalyser, har en allerede beveget seg bort fra den tradisjonelle hypotesetestingssituasjonen, hvor den alternative hypotesen er påstanden en ønsker å bevise.

Et høyere nivå gir generelt høyere styrke dersom vi holder oss til den samme testmetoden og ikke endrer sampelstørrelsen. Det finnes ingen garantier for at en annen testmetode med høyere nivå vil gi bedre styrke. Jeg ønsker likevel å teste denne muligheten. Jeg foreslår en mulig forbedring til Begg og Mazumdars testprosedyre. Denne nye testmetoden tilsvarer modellen introdusert av Begg og Mazumdar, men effektestimatene standardiseres ikke. Med andre ord nytter jeg en rangkorrelasjonstest basert på Kendalls tau, hvor t_i og v_i korreleres for $i = 1, 2, \ldots, k$. Dette er testmodellen Begg og Mazumdar argumenterer for at ikke er gyldig [8]. Jeg kaller denne testprosedyren for den ustandardiserte testen. Videre vil jeg også kalle Begg og Mazumdars testprosedyre for den standardiserte testen.

I en metaanalysesammenheng er jeg villig til å "ofre" feil av type I dersom dette bidrar til lavere feil av type II. Dette betyr ikke at en ørliten økning i styrke veier opp for en stor økning

i nivået, men utdyper at konsekvensene av en feil av type II er mer alvorlig enn en tilsvarende feil av type I. Jeg ønsker testmetoder hvor signifikansnivået er korrekt. Om ikke nivået lar seg tilpasse, foretrekkes tester med et nivå som er høyere enn det nominelle dersom dette gir tilsvarende bedre styrke. Beslutningsteori er avgjørende for argumentasjonen her. Det samme er bevisstheten om at vi ikke er i en tradisjonell hypotesetestingssituasjon.

En vil aldri kunne gjøre feil av type II dersom sannsynligheten for å forkaste H_0 er lik 1. Selv om en ikke er i en tradisjonell hypotesetestingssituasjon, må en fortsatt foreta en rimelig avveining mellom ønskene om lav feil av type I og type II.

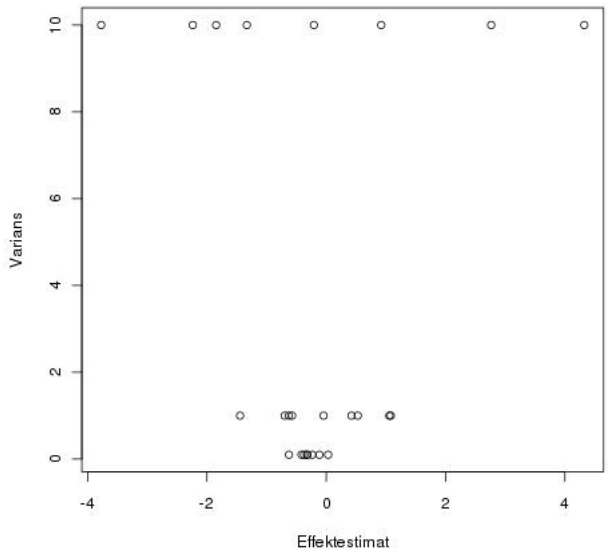
En umiddelbar fordel med den ustandardiserte testen er at den er enklere å utføre. Noen vil kanskje innvende at det ikke er poeng i å teste ut den ustandardiserte testmetoden. Effektestimatene er ikke uavhengige av variansene under nullhypotesen. Forutsetningene for å utføre rangkorrelasjonstesten basert på Kendalls tau er brutt. Begg og Mazumdar [8] har dessuten poengtert at nivået ikke er korrekt ved simuleringer av denne testprosedyren. Allerede før jeg går i gang med simuleringer, er jeg fullt klar over at den nye testprosedyren ikke vil fungere optimalt. Likevel håper jeg at den nye testprosedyren vil være en forbedring av den standardiserte testen. Simuleringer vil vise hvordan den ustandardiserte testen vil prestere. Først vil jeg gjennomgå forutsetningene som brytes når en utfører den nye testmetoden under nullhypotesen.

7.3 Forutsetninger som brytes i den ustandardiserte testen under nullhypotesen

Som tidligere nevnt er det en konvensjonell oppfatning at de k stokastiske parene (t_1, v_1), $(t_2, v_2), \ldots, (t_k, v_k)$ fra studiene i en metaanalyse er uavhengige og har den samme kontinuerlige bivariate fordelingen. Jeg antar at disse forutsetningene er oppfylt i reelle metaanalysesituasjoner.

Begg og Mazumdar lar variansene formelt sett være faste størrelser. Med faste varianser kan vi, ved hjelp av Kendalls tau, teste nullhypotesen om at $t_1, t_2, \ldots, t_k$ er uavhengige og identisk fordelte. Simuleringer i Kapittel 6, samt likheten med Mann-Kendall trendtest, understøtter at dette ikke bør by på problemer.

Effektestimatene er ikke identisk fordelte gitt variansene. Dette er kjente problemer og grunnen til at Begg og Mazumdar standardiserer effektestimatene i utgangspunktet. Fordelingen til t_i gitt $v_1, v_2, \ldots, v_k$ er normal med forventning δ og varians v_i i testsituasjonen til Begg og Mazumdar. Effektestimatene er ikke uavhengige av variansene under nullhypotesen. Dette understøttes i Figur 4. Her plottes variansene, $k = 25$ og $v = 0.1, 1.0, 10.0$, mot de tilhørende effektestimatene, t_i, hvor $\delta = 0$. Simuleringene fra Kapittel 6.3 viser hvordan nivået eksempelvis kan påvirkes dersom t_i og v_i ikke er uavhengige under nullhypotesen, $i = 1, 2, \ldots, k$.



Figur 4: Variansene plottes mot effektestimatene.

7.4 Simuleringsresultater og vurdering av den ustandardiserte testprosedyren

Sannsynligvis vil selektiv publikasjon være påvirket av både effektstørrelsen og p-verdien for hypotesen om at den sanne, underliggende effekten er lik null [5]. Tidligere simuleringsresultater viser at valg av seleksjonsmodell hovedsaklig påvirker styrken til den standardiserte testen i liten grad. Jeg vurderer testmetodens egenskaper nokså likt for begge valg av vektfunksjoner, selvfølgelig med noen unntak. Denne oppgaven handler ikke først og fremst om seleksjonsmodeller. Jeg ønsker ikke å utdype denne tematikken nærmere. De videre simuleringene avgrenses til seleksjonsfunksjonen som avhenger av p-verdien. Dette understøttes av Macaskill et al. [8], som også kun selekterer studier ved hjelp av denne vektfunksjonen.

Jeg gjentar simuleringsprosedyren beskrevet i Kapittel 3 for de ulike konfigurasjonene ved ensidig seleksjon basert på p-verdi, men utfører denne gangen den ustandardiserte testmetoden. Simuleringensprosessen repeteres 5000 ganger. Jeg velger stadig et tosidig nominelt nivå på 5%. Resultatene for metaanalysene uten publikasjonsbias finnes i Tabell 11 og Tabell 12, mens resultatene for metaanalysene med publikasjonsbias finnes i Tabell 13 og Tabell 14. Resultatene er basert på funksjonen `cor.test()` i statistikkprogrammet R [43]. Den logiske indikatoren `exact` settes lik `NULL`. For $k = 25$ beregnes en eksakt p-verdi for hypotesen om ingen publikasjonsbias. Kendalls tau er asymptotisk normalfordelt, og denne fordelingen nyttes

ikke for $k < 50$.

Jeg gjentar også simuleringsprosedyren gitt av Begg og Mazumdar ved ensidig seleksjon basert på p-verdi for den standardiserte testen. Resultatene i denne seksjonen er, i motsetning til resultatene i Kapittel 4, beregnet ved hjelp av funksjonen `cor.test()`, hvor den logiske indikatoren `exact` settes lik `NULL`. Jeg nytter ikke lenger den asymptotiske fordelingen til tau for $k = 25$. Ellers er ingen endringer foretatt sammenliknet med simuleringene utført i Kapittel 4. De nye resultatene inkluderes for å kunne vurdere den standardiserte og ustandardiserte testen opp mot hverandre på en mest mulig rettferdig måte. Simuleringsresultatene for metaanalysene uten publikasjonsbias finnes i Tabell 15 og Tabell 16. Resultatene for metaanalysene med publikasjonsbias finnes i Tabell 17 og Tabell 18. Simuleringene for den standardiserte og ustandardiserte testmetoden er utført uavhengig av hverandre.

Simuleringsresultatene bekrefter at nivåproblemene med Begg og Mazumdars testmetode eksisterer uavhengig om p-verdiene baseres på den eksakte eller asymptotiske fordelingen til Kendalls tau. Resultater eller konklusjoner påvirkes ikke vesentlig av at jeg velger å bruke den eksakte fordelingen til tau når $k = 25$.

Testmetoden, hvor de ustandardiserte effektestimatene korreleres mot variansene, viser mange av de samme tendensene som testprosedyren til Begg og Mazumdar. Styrken er betraktelig bedre for store metaanalyser enn for små. Både variansspredning og seleksjonsstyrke har stor innvirkning på styrken. Best styrke oppnås ved sterk seleksjonsbias og stor variansspredning. Styrken blir lavere jo lenger bort fra nullverdien δ beveger seg.

Signifikansivået til den ustandardiserte testen ligger omkring 0.05 eller høyere. Nivået til den standardiserte testen ligger lavere enn det nominelle for de fleste simuleringskonfigurasjoner. Tosidige konfidensintervaller med konfidensnivå 0.95 bekrefter dette. Jeg vil kort bemerke at de gjennomsnittlige nivåestimatene for den ustandardiserte testen avviker fra de rapporterte vediene til Begg og Mazumdar [8]. Forfatterne viser til et gjennomsnittlig nivåestimat på 0.10 ved stor variansspredning for $k = 25$. Det nominelle nivået skal være 0.05. Jeg estimerer nivået til å ligge omkring 0.08. Ved liten variansspredning gir de et gjennomsnittlig estimat på 0.09, også her for $k = 25$. Jeg estimerer signifikansnivået til omkring 0.05. Jeg har ingen forklaring på disse avvikene.

Det er vanskelig å gjennomføre en rettferdig sammenlikning av de to testmetodene ut fra resultatene i tabellene. Jeg klarer ikke å kontrollere feil av type I på et gitt nivå. Det er ikke lett å si om en testobservator inneholder mer informasjon enn den andre. Dette er særlig tilfellet ved stor variansspredning, hvor nivåestimatene for den ustandardiserte testen avviker stort fra den standardiserte.

Beslutningsteori er viktig for å avgjøre hvilken test en bør velge. Det er utfordrende å definere en passende tapsfunksjon. En god tapsfunksjon bør, i en metaanalysesammenheng,

også avhenge av andelen inkluderte studier og bias. En feil av type II er alvorlig dersom selektiv publikasjon medfører resultater som er biased. De to feiltypene bør vektes ulikt, men hvor mye større bør kostnaden være av å gjøre en feil av type II enn en feil av type I? En feil av type II er mindre alvorlig dersom biasen, $\beta = \mathrm{E}(\bar{t} - \delta)$, er liten. Fra et praktisk ståsted kan en argumentere at en feil av type I er mer alvorlig enn en feil av type II for store verdier av δ når biasen er liten. Tapsfunksjonen bør reflektere dette. Fordi vi ikke vet hvor stor biasen vil være i reelle metaanalyser, vil det totalt sett likevel være viktigere å begrense feil av type II enn feil av type I. Selv om jeg velger ikke å definere tapsfunksjoner og derfor heller ikke risikofunksjoner, benytter jeg i stor grad den grunnleggende tankegangen bak beslutningsteori. Ut fra denne tankegangen, gjør jeg en rimelig vurdering basert på skjønn.

Den ustandardiserte testen viser en signifikant forbedring i styrke i situasjoner hvor variansspredningen er stor, sammenliknet med den standardiserte testen. Eksempelvis estimeres styrken til 81% for den ustandardiserte testen når $k = 25$, $\delta = 0$ og seleksjonsstyrken sterk. Det tilsvarende estimatet for den standardiserte testen er 57%. Dessverre må styrken vurderes i sammenheng med nivået. Når $k = 25$ og $\delta = 0$, estimeres nivået til 0.0838 for den ustandardiserte testen. For den standardiserte testen er nivåestimatet 0.0172. Jeg tar de uvanlige rollene til hypotesene H_0 og H_1 i betraktning. I tilfeller med stor variansspredning foretrekkes den ustandardiserte testen. Denne konklusjonen gjelder både når $k = 25$ og $k = 75$ og ved moderat og sterk seleksjonsstyrke.

En forkaster ikke nullhypotesen om at signifikansnivået er lik det nominelle for den ustandardiserte testmetoden for samtlige verdier av δ når $k = 25$ og variansspredning er liten. Slutningene er basert på tosidige konfidensintervaller med konfidensnivå 0.95. Den standardiserte testen har hovedsaklig et nivå som ligger noe lavere enn 0.05. Testprosedyrene presterer nokså likt med tanke på styrke, selv om den ustandardiserte testen i enkelte tilfeller ser ut til å ha lavere feil av type II.

Jeg vurderer nå testene opp mot hverandre når variansspredningen er liten og $k = 75$. For den ustandardiserte testen forkaster vi nullhypotesen om at nivået er lik 0.05 for fire av syv verdier av δ. I forkastningstilfellene er nivået for høyt. For den standardiserte testen forkaster vi den samme nullhypotesen for fem av syv verdier av δ. Nivået ligger generelt lavere enn 0.05.

Styrken er hovedsaklig nokså lik for de to testmetodene når variansspredningen er liten og $k = 75$. Den ustandardiserte testmodellen har i enkelte tilfeller bedre styrke når seleksjonsstyrken er sterk, basert på tosidige konfidensintervaller med konfidensnivå 0.95. Hva når seleksjonsstyrken er moderat? Tabellene viser at styrkeestimatene for den ustandardiserte testen er lavere enn styrkeestimatene for den standardiserte testen for δ lik 0.0, 0.5 og 1.0. For disse verdiene av δ er biasen høy, hvilket betyr lavere styrke i tilfeller hvor betydningen av publikasjonsbias er stor. I disse tilfellene inneholder den standardiserte testobservatoren mest

informasjon. Det kan se ut til at en risikerer større feil av både type I og type II ved å utføre den ustandardiserte testen. For disse tre deltaverdiene forkaster en derimot kun hypotesen om at $p_1 = p_2$ når $\delta = 0.5$, hvor p_1 og p_2 betegner styrken for den standardiserte og ustandardiserte testmetoden henholdsvis. Når $\delta = 0.5$, estimeres styrken til 0.3124 for den standardiserte testen. Det tilsvarende estimatet for den ustandardiserte testmodellen er 0.2924. Tabell 19 viser konfidensintervallene for $p_1 - p_2$ når $k = 75$, variansspredningen liten og seleksjonsstyrken moderat. I denne tabellen nytter jeg estimatene slik de estimeres i R, uten å runde dem av slik jeg tidligere har gjort. Tabell 19 viser videre at vi bør forkaste hypotesen om at styrken er lik for de fire største verdiene av δ. Denne gangen forkastes nullhypotesen i favør av den ustandardiserte testen.

Jeg foretar en totalvurdering av resultatene med liten variansspredning, både for sterk og moderat seleksjonsstyrke og for $k = 25$ og $k = 75$ samlet. Jeg vurderer styrken i sammenheng med nivået. Hovedsaklig vil valg av test trolig ikke påvirke konklusjonene i betydelig grad.

Vi ønsker styrkerette tester. Casella og Berger [10] definerer en styrkefunksjon $\beta(\theta)$ som styrkerett dersom $\beta(\theta') \geq \beta(\theta'')$ for alle $\theta' \in \Theta_0^c$ og $\theta'' \in \Theta_0$. Styrkefunksjonen avhenger av flere parametre. Jeg klarer ikke å konstruere analytiske styrkefunksjoner for de aktuelle testsituasjonene og vil ikke kunne fastslå sikkert om testene er styrkerette. Likevel vil jeg kunne danne meg et bilde av hvordan styrkefunksjonen ser ut i forskjellige tilfeller ved hjelp av Monte Carlo-eksperimenter. Jeg har allerede testet for moderat og sterk seleksjonsstyrke ved ensidig seleksjon basert på p-verdi. En forkaster H_0 med større eller lik sannsynlighet hvis $\theta \in \Theta_0^c$ enn dersom $\theta \in \Theta_0$ for samtlige estimater for begge testemetodene.

For å kunne konkludere med større grad av sikkerhet, estimerer jeg styrken til testene på samme måte som før. Jeg letter på seleksjonspresset ved å øke verdien til a i vektfunksjonen, mens b som vanlig har verdien 4. Verdien $a = 10$ og $a = 50$ tilsvarer henholdsvis liten og svært liten seleksjonsstyrke. For $a = 10$ viser simuleringsresultatene at minimum 83% av studiene inkluderes i metaanalysen. For de fleste verdier av δ ligger denne prosenten over 95. Når $a = 50$, blir minimum 96% av studiene inkludert. De fleste estimatene ligger over 99,5%. Styrkeestimatene understøtter at testmetodene er styrkerette. Resultatene inkluderes ikke i oppgaven.

Min generelle vurdering, på bakgrunn av beslutningsteori, er at en bør velge den ustandardiserte testen. Dette er særlig tilfellet ved stor variansspredning. Ved liten variansspredning vil valg av test trolig ikke påvirke utfallet i særlig grad. Den ustandardiserte testen presterer på det jevne bedre enn Begg og Mazumdars testprosedyre, sett i en metaanalysesammenheng. En minsker stort sett sannsynligheten for å gjøre en feil av type II, dessverre på bekostning av feil av type I.

Tabell 11: Ustandardisert effektestimat korreleres mot varians. Nivå. Liten metaanalyse ($k = 25$).

	Nivå [% inkluderte studier, bias]	
Variansspredning	Stor	Liten
Behandlingseffekt (δ)		
.0	8,38%	5,26%
	[100%, -.00]	[100%, .00]
.5	8,78%	4,98%
	[100%, .00]	[100%, -.00]
1.0	8,04%	5,10%
	[100%, -.00]	[100%, -.00]
1.5	8,16%	5,38%
	[100%, .00]	[100%, -.00]
2.0	7,92%	5,20%
	[100%, -.00]	[100%, -.00]
2.5	7,86%	5,42%
	[100%, -.00]	[100%, .00]
3.0	8,38%	4,92%
	[100%, .00]	[100%, .00]

Tabell 12: Ustandardisert effektestimat korreleres mot varians. Nivå. Stor metaanalyse ($k = 75$).

	Nivå [% inkluderte studier, bias]	
Variansspredning	Stor	Liten
Behandlingseffekt (δ)		
.0	9,00%	5,64%
	[100%, .00]	[100%, -.00]
.5	9,04%	5,84%
	[100%, .00]	[100%, -.00]
1.0	9,16%	5,00%
	[100%, .00]	[100%, .00]
1.5	9,30%	5,42%
	[100%, -.00]	[100%, .00]
2.0	9,22%	5,82%
	[100%, .00]	[100%, .00]
2.5	8,68%	4,90%
	[100%, -.00]	[100%, -.00]
3.0	9,06%	6,02%
	[100%, -.00]	[100%, -.00]

Tabell 13: Ustandardisert effektestimat korreleres mot varians. Styrke ved ensidig seleksjon basert på p-verdi. Liten metaanalyse ($k = 25$).

	Styrke [% inkluderte studier, bias]			
Seleksjonsstyrke	Sterk		Moderat	
Variansspredning	Stor	Liten	Stor	Liten
Behandlingseffekt (δ)				
.0	81%	24%	57%	13%
	[36%, .34]	[37%, .74]	[57%, .25]	[57%, .54]
.5	79%	23%	49%	11%
	[54%, .16]	[52%, .55]	[74%, .09]	[73%, .35]
1.0	70%	18%	34%	10%
	[65%, .07]	[67%, .36]	[82%, .04]	[85%, .20]
1.5	57%	14%	24%	7%
	[72%, .05]	[79%, .23]	[87%, .02]	[92%, .10]
2.0	43%	9%	17%	6%
	[78%, .03]	[88%, .13]	[90%, .01]	[96%, .05]
2.5	34%	6%	15%	5%
	[82%, .03]	[93%, .07]	[92%, .01]	[98%, .03]
3.0	24%	6%	11%	6%
	[86%, .02]	[96%, .04]	[94%, .01]	[99%, .01]

Tabell 14: Ustandardisert effektestimat korreleres mot varians. Styrke ved ensidig seleksjon basert på p-verdi. Stor metaanalyse ($k = 75$).

	Styrke [% inkluderte studier, bias]			
Seleksjonsstyrke	Sterk		Moderat	
Variansspredning	Stor	Liten	Stor	Liten
Behandlingseffekt (δ)				
.0	100%	64%	96%	36%
	[36%, .33]	[36%, .74]	[56%, .24]	[56%, .54]
.5	100%	61%	91%	29%
	[53%, .16]	[52%, .54]	[74%, .08]	[72%, .34]
1.0	99%	51%	77%	20%
	[64%, .07]	[67%, .36]	[82%, .04]	[84%, .20]
1.5	96%	39%	58%	14%
	[71%, .04]	[79%, .23]	[86%, .02]	[92%, .10]
2.0	89%	24%	42%	8%
	[77%, .03]	[88%, .13]	[90%, .02]	[96%, .05]
2.5	78%	14%	31%	7%
	[81%, .02]	[93%, .08]	[92%, .01]	[98%, .02]
3.0	65%	9%	23%	6%
	[85%, .02]	[96%, .04]	[94%, .01]	[99%, .01]

Tabell 15: Standardisert effektestimat korreleres mot varians. Nivå. Liten metaanalyse ($k = 25$).

	Nivå [% inkluderte studier, bias]	
Variansspredning	Stor	Liten
Behandlingseffekt (δ)		
.0	1,72%	3,96%
	[100%, .00]	[100%, .00]
.5	1,82%	4,36%
	[100%, .00]	[100%, .00]
1.0	1,86%	4,30%
	[100%, .00]	[100%, .00]
1.5	1,90%	3,68%
	[100%, .00]	[100%, .00]
2.0	1,82%	3,58%
	[100%, .00]	[100%, -.00]
2.5	1,54%	4,48%
	[100%, .00]	[100%, .00]
3.0	1,74%	4,24%
	[100%, .00]	[100%, -.00]

Tabell 16: Standardisert effektestimat korreleres mot varians. Nivå. Stor metaanalyse ($k = 75$).

	Nivå [% inkluderte studier, bias]	
Variansspredning	Stor	Liten
Behandlingseffekt (δ)		
.0	1,76%	4,12%
	[100%, -.00]	[100%, .00]
.5	1,70%	4,74%
	[100%, .00]	[100%, -.00]
1.0	2,38%	4,54%
	[100%, .00]	[100%, .00]
1.5	1,96%	4,30%
	[100%, .00]	[100%, -.00]
2.0	1,60%	4,24%
	[100%, .00]	[100%, -.00]
2.5	1,88%	4,10%
	[100%, .00]	[100%, .00]
3.0	1,64%	4,22%
	[100%, .00]	[100%, .00]

Tabell 17: Standardisert effektestimat korreleres mot varians. Styrke for ensidig seleksjon basert på p-verdi. Liten metaanalyse ($k = 25$).

	Styrke [% inkluderte studier, bias]			
Seleksjonsstyrke	Sterk		Moderat	
Variansspredning	Stor	Liten	Stor	Liten
Behandlingseffekt (δ)				
.0	57%	22%	33%	13%
	[36%, .34]	[37%, .74]	[57%, .25]	[57%, .54]
.5	51%	21%	23%	11%
	[54%, .16]	[52%, .54]	[74%, .09]	[73%, .34]
1.0	39%	16%	13%	8%
	[65%, .07]	[67%, .36]	[82%, .04]	[85%, .20]
1.5	27%	13%	9%	6%
	[72%, .05]	[80%, .23]	[87%, .02]	[92%, .10]
2.0	19%	8%	5%	5%
	[78%, .03]	[88%, .14]	[90%, .02]	[96%, .05]
2.5	12%	6%	3%	4%
	[82%, .02]	[93%, .07]	[93%, .01]	[98%, .03]
3.0	9%	5%	3%	4%
	[86%, .02]	[96%, .04]	[94%, .01]	[99%, .01]

Tabell 18: Standardisert effektestimat korreleres mot varians. Styrke for ensidig seleksjon basert på p-verdi. Stor metaanalyse ($k = 75$).

	Styrke [% inkluderte studier, bias]			
Seleksjonsstyrke	Sterk		Moderat	
Variansspredning	Stor	Liten	Stor	Liten
Behandlingseffekt (δ)				
.0	99%	61%	88%	38%
	[36%, .34]	[36%, .74]	[56%, .24]	[56%, .54]
.5	99%	59%	77%	31%
	[53%, .16]	[52%, .54]	[74%, .09]	[72%, .34]
1.0	94%	50%	54%	21%
	[64%, .07]	[67%, .36]	[82%, .04]	[84%, .19]
1.5	85%	35%	35%	12%
	[71%, .04]	[79%, .23]	[86%, .02]	[92%, .10]
2.0	71%	22%	21%	7%
	[77%, .03]	[88%, .13]	[90%, .02]	[96%, .05]
2.5	53%	12%	13%	5%
	[81%, .02]	[93%, .07]	[92%, .01]	[98%, .03]
3.0	40%	7%	8%	5%
	[85%, .02]	[96%, .04]	[94%, .01]	[99%, .01]

Tabell 19: Styrkeestimat og konfidensintervall for stor metaanalyse, liten variansspredning.

Styrkeestimat, standardisert test	Styrkeestimat, ustandardisert test	95% konfidensintervall for $p_1 - p_2$
0.3780	0.3612	[-0,00212, 0.03572]
0.3124	0.2924	[0.00200, 0.03800]
0.2120	0.2008	[-0.00466, 0.02706]
0.1226	0.1386	[-0.02921, -0.00279]
0.0658	0.0822	[-0.02666, -0.00614]
0.0458	0.0662	[-0.02940, -0.01140]
0.0466	0.0560	[-0.01805, -0.00075]

Styrken for den standardiserte og ustandardiserte testmetoden betegnes henholdsvis p_1 og p_2.

8 Forslag til forbedring av testmetoder basert på den simulerte fordelingen til Kendalls tau

Fra kapittel 5.2 vet vi at sannsynlighetsfordelingen til Kendalls tau i Begg og Mazumdars testmodell ikke er lik sannsynlighetsfordelingen til den ordinære Kendalls tau. Optimalt ønsker jeg en test basert på rangkorrelasjoner hvor en lett kan tilpasse nivået. Dette kan oppnås ved å nytte den simulerte fordelingen til det estimerte assosiasjonsmålet. Vi betinger på de estimerte variansene. Da kan vi se bort fra de forutsetningene som brytes under nullhypotesen. Signifikansnivået vil være tilnærmet lik det nominelle.

8.1 Beskrivelse av testmetoder med tilnærmet korrekt nivå

Det skal gjennomføres en rangkorrelasjonstest for å identifisere publikasjonbias i en metaanalyse. Vi har tilgjengelig k effektestimater, $t_1, t_2, \ldots, t_k$, med tilhørende varianser $v_1, v_2, \ldots, v_k$. Jeg ønsker å tilpasse nivået til rangkorrelasjonstesten basert på Kendalls tau ved å nytte en form for parametrisk bootstrapping. I første omgang ser jeg bort fra de tilgjengelige effektestimatene i metaanalysen. Ut fra de k tilgjengelige variansene, genereres k *nye* normalfordelte effektestimater, $\tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_k$. Variansen til $\tilde{t}_i$ er altså v_i. Uten tap av generalitet settes forventningen til $\tilde{t}_i$ lik null. Nivåestimatene ser ikke ut til å avhenge av den underliggende behandlingseffekten dersom effektestimatene er normalfordelte. Dernest utføres rangkorrelasjonstesten, hvor $\tilde{t}$ og v korreleres. Den beregnede verdien til Kendalls tau lagres. Prosedyren repeteres n antall ganger, slik at jeg har n reproduserte verdier av tau, $\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_n$. Jeg skulle fortrinnsvis nyttet notasjonen t. I Kapittel 2 er τ parameterverdien til Kendalls tau, t. For å unngå forvirring mellom notasjonen til effektestimatene og Kendalls tau, betegner $\hat{\tau}$ estimatoren til parameteren τ. Den empiriske fordelingsfunksjonen til de simulerte verdiene til Kendalls tau er gitt ved

$$F_n(\hat{\tau}) = \begin{cases} 0 & \text{hvis } \hat{\tau} < \hat{\tau}_{(1)} \\ \dfrac{i}{n} & \text{hvis } \hat{\tau}_{(i)} \leq \hat{\tau} < \hat{\tau}_{(i+1)}, \quad i = 1, 2, \ldots, n-1 \\ 1 & \text{hvis } \hat{\tau}_{(n)} \leq \hat{\tau} \end{cases} .$$

Her er $\hat{\tau}_{(1)} \leq \hat{\tau}_{(2)} \leq \cdots \leq \hat{\tau}_{(n)}$ de ordnede verdiene til $\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_n$ [46].

Jeg ønsker videre å finne forkastningsintervallene for Kendalls tau i denne aktuelle situasjonen. Det vil si jeg ønsker å finne de verdiene av tau som gir forkastning av nullhypotesen om ingen publikasjonsbias, slik at det virkelige nivået er lik det nominelle. Forkastningsintervallene kan finnes ved ulike framgangsmåter. Eksempelvis kan en konstruere konfidensintervaller basert på antakelsen om at estimatoren er tilnærmet normalfordelt. Vi antar at $\hat{\tau}$ er en estimator for parameteren τ. Videre antar vi at $\text{se}(\hat{\tau})$ er standardfeilen til denne estimatoren. Hvis $\hat{\tau}$

er normalfordelt eller et gjennomsnitt hvor sampelstørrelsen er stor, gir sentralgrenseteoremet at $z = \big(\hat{\tau} - \mathrm{E}(\hat{\tau})\big)/\mathrm{se}(\hat{\tau})$ er tilnærmet standardnormalfordelt. Dersom $\hat{\tau}$ er en forventningsrett estimator for τ, vil et tilnærmet konfidensintervall for τ med konfidensnivå $1-\alpha$ være gitt ved $\hat{\tau} \pm z_{\alpha/2}\mathrm{se}(\hat{\tau})$, hvor $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ [46]. Konfidensintervallet er enkelt å beregne, men kan være unøyaktig grunnet de overnevnte antakelsene. Vi har behandlet $\mathrm{se}(\hat{\tau})$ som en kjent parameter. I virkeligheten må standardfeilen estimeres.

Alternativt kan en finne de ønskede kvantilene, $\alpha/2$- og $(1-\alpha/2)$-kvantilene, til den empiriske fordelingen til de simulerte verdiene av Kendalls tau. Kvantilene til den empiriske fordelingen er estimatorer for kvantilene til sampelfordelingen til $\hat{\tau}$ [46]. Denne metoden er ikke-parametrisk. Den skal ha visse teoretiske og praktiske fordeler framfor metoden basert på antakelsen om normalfordelte estimatorer. Likevel kan også kvantilmetoden modifiseres for å oppnå bedre teoretiske egenskaper og bedre prestasjoner i praksis [46]. Jeg vil ikke gå nærmere inn på ulike metoder for å finne konfidensintervaller, men velger å holde meg til kvantilmetoden. Denne er intuitiv og lett å forstå. For mer lettlest informasjon henvises leseren til Rizzo [46].

Forkastningsintervallene vil være estimert med feil. Disse avvikene vil forplante seg videre og kan gi alvorlige konsekvenser når en senere skal teste om en metaanalyse inneholder publikasonsbias. Jo flere ganger prosedyren gjentas, desto mindre vil denne feilen bli. En bør velge n så stor at en stoler på estimatene.

Når forkastningsintervallene er funnet, beregnes $\hat{\tau}$ ved å korrelere effektestimatene fra den aktuelle metaanalysen, $t_1, t_2, \ldots, t_k$, mot de tilhørende variansene $v_1, v_2, \ldots, v_k$. Dersom korrelasjonskoeffisienten ligger i forkastningsintervallene, forkaster vi nullhypotesen om ingen publikasjonsbias. Vi forkaster ikke denne nullhypotesen dersom $\hat{\tau}$ ikke ligger i forkastningsområdet.

Metoden kan oppsummeres i en enkel algoritme:

Gitt k effektestimater, $t_1, t_2, \ldots, t_k$, og de tilhørende variansene, $v_1, v_2, \ldots, v_k$.

1. For hver gjentakelse av prosedyren, indeksert ved $j = 1, 2, \ldots, n$:
   (a) Generer k effektestimater $\tilde{t}_1^{(j)}, \tilde{t}_2^{(j)}, \ldots, \tilde{t}_k^{(j)}$, hvor $\tilde{t}_i^{(j)}$ er normalfordelt med forventning lik null og varians v_i.
   (b) Korreler $\tilde{t}_1^{(j)}, \tilde{t}_2^{(j)}, \ldots, \tilde{t}_k^{(j)}$ og $v_1, v_2, \ldots, v_k$ for å finne korrelasjonskoeffisienten $\hat{\tau}_j$.
2. Finn forkastningsintervallene (eksempelvis ved å finne de ønskede kvantilene fra den empiriske fordelingen til $\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_n$).
3. Beregn $\hat{\tau}$ ved å korrelere $t_1, t_2, \ldots, t_k$ og $v_1, v_2, \ldots, v_k$.

4. Forkast nullhypotesen om ingen publikasjonsbias dersom $\hat{\tau}$ er i forkastningsområdet.

Vedlegg D presenterer programkode som viser hvordan algoritmen kan implementeres i praksis. Jeg ønsker også å estimere den aktuelle tosidige mid-p-verdien. Programkoden viser hvordan dette kan gjøres. Mid-p likner definisjonen av vanlig p-verdi, med unntak av at en kun inkluderer halve sannsynligheten av akkurat den verdien man har registrert av testobservatoren. Mid-p erstatter den vanlige p-verdien fordi signifikanstestene er basert på diskrete sannsynlighetsfordelinger. Den interesserte leser henvises til Lancaster [29] for videre teori om mid-p.

Dersom fordelingen til $\hat{\tau}$ er asymmetrisk, er det ikke rett fram å finne tosidig p-verdi. Det finnes ingen generelt akseptert metode for hvordan dette bør gjøres [3]. Likevel finner jeg en tilnærmet tosidig p-verdi ved å doble den ensidige. Denne metoden kan forsvares ved at en asymmetrisk kontinuerlig fordeling med et enkelt maksimumspunkt alltid kan transformeres til en normalfordeling [57]. En tosidig test med nivå α avsetter halve verdien av α til å teste for statistisk signifikans i en retning. Den andre halvdelen avsettes til å teste for statistisk signifikans i den motsatte retningen. En dobling av den ensidige p-verdien samsvarer med denne tankegangen. Selv om en også kan argumentere mot denne prosedyren, vil metoden fungere bra i tilnærmede symmetriske situasjoner.

Jeg har gjennomført punkt 1 i algoritmen ovenfor for ulike valg av varianser som en typisk kan finne i reelle metaanalyser. Histogrammer viser at det ikke er urimelig å anta at $\hat{\tau}$ er normalfordelt. En dobling av den ensidige mid-p-verdien vil være en grei tilnærming til den tosidige mid-p-verdien. Beregning av skjevhet kan også gi nyttig informasjon. En kan undersøke om den empiriske fordelingen er normalfordelt ved hjelp av en test basert på denne skjevheten [46]. Dette prioriteres ikke i denne oppgaven. Jeg antar videre at den empiriske fordelingen er symmetrisk om medianen.

Testmetoden jeg har beskrevet kan brukes med utgangspunkt i både den standardiserte og den ustandardiserte testobservatoren. Algoritmen ovenfor brukes direkte dersom vi tar utgangspunkt i den ustandardiserte observatoren. Jeg vil kalle dette for den korrigerte ustandardiserte testmetoden, KUT. Algoritmen lar seg lett generalisere om en ønsker å ta utgangspunkt i den standardiserte testobservatoren. Punkt 1(a) i algoritmen vil være den samme, men de nye genererte effektestimatene må standardiseres. Deretter korreleres de *nye* standardiserte effektestimatene og variansene. Jeg kaller dette for den korrigerte standardiserte testmetoden, KST. De korrigerte testmetodene sikter helt enkelt til framgangsmåten hvor forkastningsintervallene for de *eksisterende* testobservatorene justeres, slik at signifikansnivået er tilnærmet lik det nominelle.

8.2 Utfordringer knyttet til de korrigerte testmetodene

Som tidligere beskrevet bygger effektestimatene på bidrag fra mange observasjoner og vil ha en asymptotisk normalfordeling. Antakelsen om normalfordeling er ikke urimelig for studier hvor sampelstørrelsen er stor. Når studienes sampelstørrelse er liten, kan man ikke uten videre anta en asymptotisk normalfordeling. Det er ønskelig å utvikle ikke-parametriske testprosedyrer. De korrigerte testmetodene har rom for utbedringer.

I de korrigerte testmodellene betinger vi på variansene i utvalget. Variansene behandles som faste størrelser. Det kan tenkes at bedre metoder muligens kan utvikles om en tar hensyn til at variansene i virkeligheten er stokastiske variable.

En annen mulig ulempe er at den empiriske fordelingen til Kendalls tau kan påvirkes dersom variansene estimeres med feil. Vi kan eksempelvis tenke oss situasjoner hvor effektestimatene og variansene estimeres ut fra felles observerte størrelser. Både effektestimatene og variansene estimeres med feil. Disse feilene kan være korrelerte. Vi risikerer skjevhet i estimatet når vi tester for publikasjonsbias. Denne problematikken utdypes nærmere i forbindelse med testmetoder basert på regresjon, se Seksjon 10.2.3.

Vil den simulerte fordelingen til Kendalls tau for disse testmodellene være robust dersom variansene systematisk underestimeres? Kendalls tau er skaleringsinvariant. Jeg forventer at systematisk underestimering av variansene ikke vil ødelegge for korrigeringsmetodene. Dette understøttes ved simuleringer i Seksjon 8.3.1.

Ingen av de overnevnte ulempene vil påvirke de korrigerte testenes prestasjoner i simuleringssituasjonene til Begg og Mazumdar, beskrevet i Kapittel 3. Effektestimatene er normalfordelte. Variansene er faste størrelser, målt uten feil. Jeg har tro på at de korrigerte testene vil rette opp i problemene knyttet til Begg og Mazumdars testmodell for disse scenarioene.

8.3 Simuleringer, resultater og sammenlikning av testobservatorenes informasjon

Jeg gjennomfører både den korrigerte standardiserte testmetoden og den korrigerte ustandardiserte testprosedyren i simuleringsscenarioene beskrevet i Kapittel 3. Igjen velger jeg kun å ta for meg ensidig seleksjon basert på p-verdi for hypotesen om at den underliggende effekten er lik null. Først beregnes forkastningsintervallene for hver av de korrigerte testmetodene. Det vil si jeg finner $\alpha/2$- og $(1-\alpha/2)$-kvantilene, hvor $\alpha = 0.05$. Disse er beregnet ved å nytte punkt 1 og 2 i algoritmen i forrige seksjon, hvor n velges lik 10000. Fordi nivået ikke ser ut til å avhenge av den underliggende behandlingseffekten i disse situasjonene, bruker jeg de samme forkastningsintervallene for samtlige verdier av δ for en gitt verdi av k i kombinasjon med en gitt variansspredning. Forkastningsintervallene beregnes ut fra den simulerte fordelingen til

Kendalls tau, ikke ut fra den eksakte eller asymptotiske fordelingen til den ordinære Kendalls tau. Nivå- og styrkeestimatene er basert på 10000 gjentakelser av simuleringsprosedyren beskrevet av Begg og Mazumdar, hvor forkastningsintervallene nå er justert.

Nivåestimatene for KST og KUT presenteres i Tabell 20-23. Signifikansnivået er tilnærmet lik 0.05 for begge testmetodene, for samtlige konfigurasjoner.

Styrkeestimatene for de korrigerte testmetodene er gitt i Tabell 24-27. Styrkeestimatene er presentert med høyere presisjon enn tidligere. Generelt er styrken til de korrigerte testene bedret sammenliknet med styrken til Begg og Mazumdars testmodell. Styrken er hovedsaklig dårligere sammenliknet med den ustandardiserte testen. Det er ikke uventet at KST har bedre styrke enn den standardiserte testmodellen, eller at KUT har lavere styrke enn den ustandardiserte testen. Et økt nivå medfører generelt økt styrke for den samme testobservatoren om sampelstørrelsen holdes konstant. Korrigeringsmetodene anbefales fordi jeg klarer å tilpasse signifikansnivået.

Siden feil av type II er mer alvorlig enn feil av type I, kan en gjerne velge et høyere signifikansnivå ved testing for publikasjonsbias. Macaskill et al. [35] velger et nominelt nivå på 0.10, et valg også Begg ser ut til å understøtte [5]. Dette er et argument for å undersøke testmetodenes virkelige nivå dersom det nominelle nivået settes til 0.10. En videre vurdering av de korrigerte testmetodene er også ønskelig i situasjoner som ligger tettere opp mot reelle meta-analyser. Jeg vil dessverre ikke gjennomføre slike simuleringer i denne oppgaven og overlater dette til videre arbeid.

8.3.1 Er den simulerte fordelingen til Kendalls tau robust dersom variansene systematisk underestimeres?

Jeg velger $k = 25$, $\delta = 0$ og $v = 0.5, 1.0, 2.0$. Et tosidig nominelt nivå settes lik 0.05. Forkastningsintervallene, 2.5- og 97.5-kvantilene, kalkuleres ved å nytte punkt 1 og 2 i algoritmen, Seksjon 8.1. Som vanlig settes n lik 10000. Ut fra disse forkastningsområdene estimeres signifikansnivået til KST og KUT for de samme parameterverdiene. Nivåestimatene er basert på 10000 gjentakelser av simuleringsprosessen beskrevet i Kapittel 3, men med et korrigert forkastningsområde. Nivået estimeres til 0.0521 for KST og 0.0475 for KUT.

Jeg dobler verdien av variansene slik at $v = 1.0, 2.0, 4.0$. Fortsatt velger jeg $\delta = 0$ og $k = 25$. Jeg estimerer nivået for disse parametervalgene ved å bruke Begg og Mazumdars simuleringsprosedyre, men nytter de samme forkastningsintervallene som jeg fant for $v = 0.5, 1.0, 2.0$. Det estimerte nivået er 0.0527 og 0.0465 for henholdsvis KST og KUT. En tidobling av variansene gir også nivåestimater omkring 0.05, fortsatt uten å beregne nye forkastningsområder basert på $v = 5.0, 10.0, 20.0$. Resultatene understøtter at den simulerte fordelingen til Kendalls tau er robust for disse testmodellene om variansene systematisk underestimeres.

8.3.2 Sammenlikning av informasjonen til den standardiserte og ustandardiserte testobservatoren

Forkastningsintervallene for den standardiserte og ustandardiserte testobservatoren korrigeres. Signifikansnivået er lik det nominelle. Testobservatorenes informasjon kan sammenliknes ved å vurdere styrken. La p_1 og p_2 være styrken for henholdsvis KST og KUT ved de samme, bestemte parameterverdiene. De estimerte verdiene for p_1 finnes i Tabell 24 og 25. Tabell 26 og 27 inneholder de estimerte verdiene for p_2. Jeg har beregnet tosidige konfidensintervaller med konfidensnivå 0.95 for $p_1 - p_2$ for samtlige konfigurasjoner. Grunnet tabellenes omfang, presenteres ikke konfidensintervallene i oppgaven.

For små metaanalyser med stor variansspredning og sterk seleksjonsstyrke forkastes nullhypotesen om at $p_1 = p_2$ for samtlige verdier av δ. Konfidensintervallene er tosidige. Jeg kan ikke slutte at den korrigerte ustandardiserte testen presterer bedre enn den korrigerte standardiserte med tanke på styrke. Det er ikke bra å konstruere ensidige tester etter tosidige. Dette kan oppfattes som juks. Konfidensintervallene gir likevel nyttig informasjon. Den ustandardiserte testobservatoren inneholder mest informasjon.

Når det gjelder små metaanalyser for øvrig, vil en i de fleste situasjoner ikke forkaste nullhypotesen om at $p_1 = p_2$. I hovedsak inneholder testobservatorene omtrent den samme mengden med informasjon. Muligens kan en i få tilfeller oppnå bedre styrke ved å nytte den korrigerte ustandardiserte testen når variansspredningen er stor og seleksjonsstyrken moderat. Den korrigerte standardiserte testen presterer generelt bedre enn den korrigerte ustandardiserte for verdier av δ nær null når variansspredningen er liten.

For store metaanalyser med stor variansspredning og sterk seleksjonsstyrke vil jeg også konkludere at testene presterer nokså likt med tanke på styrke. Denne slutningen vil jeg ikke trekke for store metaanalyser generelt. Den standardiserte testobservatoren inneholder for det meste mer informasjon enn den ustandardiserte i disse situasjonene.

Det er vanskelig å trekke klare slutninger ut fra simuleringsresultatene. En grov konklusjon er at den korrigerte ustandardiserte testmetoden presterer likt eller bedre enn den korrigerte standardiserte i små metaanalyser når variansspredningen er stor. Den ustandardiserte testobservatoren inneholder mest informasjon. I de gjenværende tilfellene er konklusjonen motsatt, men en vil ikke utelukkende oppnå mer informasjon ved å nytte den standardiserte testobservatoren.

Tabell 20: Korrigert standardisert testmetode. Nivå. Liten metaanalyse ($k = 25$).

	Nivå [% inkluderte studier, bias]	
Varianspredning	Stor	Liten
Behandlingseffekt (δ)		
.0	4.99%	5.27%
	[100%, .00]	[100%, -.01]
.5	5.10%	5.28%
	[100%, -.00]	[100%, -.00]
1.0	5.34%	5.23%
	[100%, -.00]	[100%, -.00]
1.5	5.36%	5.10%
	[100%, .00]	[100%, -.00]
2.0	4.90%	5.22%
	[100%, .00]	[100%, .00]
2.5	4.90%	5.08%
	[100%, -.00]	[100%, -.00]
3.0	4.95%	4.97%
	[100%, -.00]	[100%, .00]

Tabell 21: Korrigert standardisert testmetode. Nivå. Stor metaanalyse ($k = 75$).

	Nivå [% inkluderte studier, bias]	
Varianspredning	Stor	Liten
Behandlingseffekt (δ)		
.0	4.89%	5.28%
	[100%, .00]	[100%, -.00]
.5	4.91%	4.82%
	[100%, -.00]	[100%, -.00]
1.0	4.91%	4.82%
	[100%, .00]	[100%, .00]
1.5	4.79%	4.82%
	[100%, .00]	[100%, -.00]
2.0	5.09%	5.03%
	[100%, -.00]	[100%, -.00]
2.5	4.90%	4.98%
	[100%, .00]	[100%, .00]
3.0	4.97%	4.73%
	[100%, .00]	[100%, .00]

Tabell 22: Korrigert ustandardisert testmetode. Nivå. Liten metaanalyse ($k = 25$).

	Nivå [% inkluderte studier, bias]	
Varianspredning	Stor	Liten
Behandlingseffekt (δ)		
.0	5.04%	4.98%
	[100%, .00]	[100%, .00]
.5	5.10%	4.75%
	[100%, .00]	[100%, .00]
1.0	5.09%	4.78%
	[100%, .00]	[100%, .00]
1.5	4.81%	5.09%
	[100%, .00]	[100%, -.00]
2.0	4.87%	4.95%
	[100%, -.00]	[100%, -.00]
2.5	5.07%	5.03%
	[100%, -.00]	[100%, .00]
3.0	4.97%	5.29%
	[100%, .00]	[100%, .00]

Tabell 23: Korrigert ustandardisert testmetode. Nivå. Stor metaanalyse ($k = 75$).

	Nivå [% inkluderte studier, bias]	
Varianspredning	Stor	Liten
Behandlingseffekt (δ)		
.0	4.62%	5.14%
	[100%, -.00]	[100%, -.00]
.5	5.05%	5.29%
	[100%, -.00]	[100%, -.00]
1.0	5.03%	5.24%
	[100%, .00]	[100%, .00]
1.5	4.81%	5.38%
	[100%, .00]	[100%, .00]
2.0	4.55%	5.09%
	[100%, -.00]	[100%, -.00]
2.5	4.91%	4.96%
	[100%, -.00]	[100%, -.00]
3.0	4.99%	5.07%
	[100%, -.00]	[100%, -.00]

Tabell 24: Korrigert standardisert testmetode. Styrke for ensidig seleksjon basert på p-verdi. Liten metaanalyse ($k = 25$).

	Styrke [% inkluderte studier, bias]			
Seleksjonsstyrke	Sterk		Moderat	
Variansspredning	Stor	Liten	Stor	Liten
Behandlingseffekt (δ)				
.0	70.81%	23.94%	47.15%	15.69%
	[36%, .34]	[36%, .74]	[57%, .25]	[57%, .54]
.5	67.14%	22.55%	37.03%	13.25%
	[54%, .16]	[52%, .54]	[74%, .09]	[73%, .34]
1.0	53.66%	20.07%	23.92%	10.21%
	[65%, .07]	[67%, .36]	[82%, .04]	[85%, .20]
1.5	41.55%	14.41%	15.83%	7.30%
	[72%, .05]	[80%, .23]	[87%, .03]	[92%, .10]
2.0	31.01%	10.03%	11.17%	6.06%
	[78%, .03]	[88%, .13]	[90%, .02]	[96%, .05]
2.5	22.07%	6.80%	8.36%	5.22%
	[82%, .02]	[93%, .08]	[93%, .01]	[98%, .03]
3.0	16.39%	5.66%	7.13%	4.68%
	[86%, .02]	[96%, .04]	[94%, .01]	[99%, .01]

Tabell 25: Korrigert standardisert testmetode. Styrke for ensidig seleksjon basert på p-verdi. Stor metaanalyse ($k = 75$).

	Styrke [% inkluderte studier, bias]			
Seleksjonsstyrke	Sterk		Moderat	
Variansspredning	Stor	Liten	Stor	Liten
Behandlingseffekt (δ)				
.0	99.64%	62.18%	94.66%	40.12%
	[36%, .34]	[36%, .74]	[56%, .24]	[56%, .54]
.5	99.39%	60.68%	87.17%	33.00%
	[53%, .16]	[52%, .54]	[74%, .09]	[73%, .34]
1.0	97.34%	52.58%	69.69%	23.93%
	[64%, .07]	[67%, .36]	[82%, .04]	[84%, .20]
1.5	92.83%	39.04%	50.02%	13.09%
	[71%, .04]	[79%, .23]	[86%, .02]	[92%, .10]
2.0	83.43%	23.65%	35.42%	7.42%
	[77%, .03]	[88%, .13]	[90%, .02]	[96%, .05]
2.5	69.68%	13.42%	23.80%	5.40%
	[81%, .02]	[93%, .07]	[92%, .01]	[98%, .03]
3.0	55.49%	7.74%	16.52%	4.59%
	[85%, .02]	[96%, .04]	[94%, .01]	[99%, .01]

Tabell 26: Korrigert ustandardisert testmetode. Styrke for ensidig seleksjon basert på p-verdi. Liten metaanalyse ($k = 25$).

	Styrke [% inkluderte studier, bias]			
Seleksjonsstyrke	Sterk		Moderat	
Variansspredning	Stor	Liten	Stor	Liten
Behandlingseffekt (δ)				
.0	74.51%	22.98%	46.28%	12.50%
	[36%, .34]	[37%, .74]	[57%, .25]	[57%, .54]
.5	72.10%	20.70%	39.00%	10.33%
	[54%, .16]	[52%, .54]	[74%, .09]	[73%, .35]
1.0	60.86%	17.58%	25.93%	8.24%
	[65%, .07]	[67%, .37]	[82%, .04]	[85%, .20]
1.5	46.67%	13.46%	16.34%	6.29%
	[72%, .05]	[80%, .23]	[87%, .3]	[92%, .10]
2.0	33.30%	9.27%	12.05%	5.52%
	[78%, .03]	[88%, .13]	[90%, .01]	[96%, .05]
2.5	24.23%	6.51%	9.08%	5.11%
	[82%, .02]	[93%, .08]	[93%, .01]	[98%, .03]
3.0	18.49%	5.22%	7.29%	5.28%
	[86%, .02]	[97%, .04]	[94%, .01]	[99%, .01]

Tabell 27: Korrigert ustandardisert testmetode. Styrke for ensidig seleksjon basert på p-verdi. Stor metaanalyse ($k = 75$).

	Styrke [% inkluderte studier, bias]			
Seleksjonsstyrke	Sterk		Moderat	
Variansspredning	Stor	Liten	Stor	Liten
Behandlingseffekt (δ)				
.0	99.70%	62.28%	92.40%	33.94%
	[36%, .34]	[36%, .74]	[56%, .24]	[57%, .54]
.5	99.60%	59.28%	86.31%	28.77%
	[53%, .16]	[52%, .54]	[74%, .09]	[73%, .34]
1.0	98.46%	49.04%	67.37%	19.84%
	[64%, .07]	[67%, .36]	[82%, .04]	[85%, .20]
1.5	93.97%	36.36%	46.34%	11.83%
	[71%, .04]	[79%, .23]	[86%, .02]	[92%, .10]
2.0	82.36%	21.97%	31.86%	7.08%
	[77%, .03]	[88%, .13]	[90%, .02]	[96%, .05]
2.5	68.22%	12.50%	20.76%	5.86%
	[81%, .02]	[93%, .08]	[92%, .01]	[98%, .03]
3.0	54.23%	8.12%	15.13%	5.44%
	[85%, .02]	[96%, .04]	[94%, .01]	[99%, .01]

9 Regresjon

Til nå har jeg studert testmetoder for identifisering av publikasjonsbias i metaanalyser som alle er basert på rangkorrelasjon. I litteraturen er det utarbeidet flere testmetoder som er basert på regresjon. Disse vil jeg undersøke nærmere. Jeg starter med et teorikapittel som omhandler regresjon. Sammenliknet med den generelle litteraturen vil jeg fokusere noe ekstra på inferens om skjæringspunktet, α. Kapittelet kan ved første øyekast virke omfattende. Teorien danner bakgrunnen for testmetodene basert på regresjon og vil være sentral i etterfølgende kapitler.

9.1 Enkel lineær regresjon

I mange situasjoner ønsker en å undersøke sammenhengen mellom to variable. Ofte er den ene variabelen, x, kjent på forhånd, og en ønsker å predikere en framtidig variabel Y. Siden Y er stokastisk, kan man ikke predikere $Y = y$ nøyaktig. I stedet estimerer en forventningen til Y [22]. Man kaller ofte Y for den avhengige variabelen eller responsvariabelen. Den uavhengige variabelen er x. Et annet navn er forklaringsvariabelen. I flere tilfeller er det rimelig å anta at forventningen til Y er en lineær funksjon av x, slik at en kan nytte uttrykket $\mathrm{E}(Y) = \alpha + \beta x$. Her er α og β størrelser som karakteriserer sammenhengen mellom den uavhengige og den avhengige variabelen. Linjen $y = \alpha + \beta x$ kalles regresjonslinjen for Y med hensyn på x [32]. Jeg gjør rede for forutsetningene for å utføre en enkel, lineær regresjonsanalyse dersom en også ønsker å utføre inferens.

Variablene $Y_1, Y_2, \ldots, Y_n$ er uavhengige og normalfordelte. Vi har sammenhengen $\mathrm{E}(Y_i) = \alpha + \beta x_i$, hvor $\mathrm{Var}(Y_i) = \sigma^2$ for $i = 1, 2, \ldots, n$. Forklaringsvariablene skal være målt uten feil.

Modellen kan også formuleres ved $Y_i = \alpha + \beta x_i + \epsilon_i$, hvor ϵ_i er uavhengige feilledd. Feilleddene er normalfordelte med forventning lik null og varians σ^2. Også her skal $x_1, x_2, \ldots, x_n$ selvsagt være målt uten feil.

9.1.1 Minste kvadraters metode

Jeg ønsker å estimere regresjonslinjen og trenger informasjon om α og β. I praksis vil disse være ukjente størrelser. Informasjon kan skaffes ved å gjenta et eksperiment n ganger, hvor vi observerer responsen for en rekke ulike verdier av forklaringsvariabelen [32]. Da har vi tilgjengelig n observasjonspar, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. På bakgrunn av de observerte parene kan jeg utlede punktestimatorer for α og β. Jeg nytter en framgangsmåte kjent som minste kvadraters metode.

Jeg innfører $S = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$. Minste kvadraters estimatorer finnes ved å minimere S med hensyn på henholdsvis α og β. Jeg deriverer S med hensyn på α og setter dette utrykket

lik null, slik at

$$\frac{\partial S}{\partial \alpha} = -2\sum(y_i - \alpha - \beta x_i) = 0.$$

Deretter deriverer jeg S med hensyn på β og setter også dette utrykket lik null. Dette gir

$$\frac{\partial S}{\partial \beta} = -2\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)x_i = 0.$$

Vi kaller disse to likningene normalikningene. En punktestimator for α er gitt ved

$$\hat{\alpha} = \frac{1}{n}\sum y_i - \frac{1}{n}\hat{\beta}\sum x_i = \bar{y} - \hat{\beta}\bar{x}.$$

Den siste normallikningen kan uttrykkes ved

$$\sum y_i x_i - \alpha\sum x_i - \beta\sum x_i^2 = 0.$$

Jeg erstatter α med $\bar{y} - \beta\bar{x}$ slik at

$$\sum y_i x_i - (\bar{y} - \beta\bar{x})\sum x_i - \beta\sum x_i^2 = 0.$$

Punktestimatoren for β er gitt ved

$$\hat{\beta} = \frac{\sum y_i x_i - \bar{y}\sum x_i}{\sum x_i^2 - \sum \bar{x} x_i},$$

som ved hjelp av enkel regning kan omskrives til

$$\hat{\beta} = \frac{\sum y_i(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}.$$

Før datasettet samles inn, kan vi oppfatte regresjonslinjen som stokastisk, lik $Y = \alpha + \beta x$. Da får vi

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x},$$

og

$$\hat{\beta} = \frac{\sum Y_i(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}.$$

Minimert kvadratsum er

$$\hat{S} = \sum(Y_i - \hat{Y}_i)^2 = \sum(Y_i - \hat{\alpha} - \hat{\beta}x_i)^2.$$

9.1.2 Utledning av testobservatorer

Jeg ønsker å utføre inferens om α og β. Av den grunn må jeg utlede testobservatorer. Jeg tar utgangspunkt i et bevis hentet fra forelesningsnotater gitt av Ivar Heuch [20]. Vi innfører $t_i = x_i - \bar{x}$, slik at $\sum t_i = 0$. Da har vi at

$$\mathrm{E}(Y_i) = \alpha + \beta x_i = \alpha + \beta(t_i + \bar{x}) = \alpha + \beta\bar{x} + \beta t_i.$$

Jeg definerer $\alpha_1 = \alpha + \beta\bar{x}$ og får $\mathrm{E}(Y_i) = \alpha_1 + \beta t_i$. Med denne notasjonen er

$$\hat{\alpha}_1 = \hat{\alpha} + \hat{\beta}\bar{x} = \bar{Y} - \hat{\beta}\bar{x} + \hat{\beta}\bar{x} = \bar{Y}$$

og

$$\hat{\beta} = \frac{\sum Y_i t_i}{\sum t_i^2}.$$

Minimert kvadratsum kan uttrykkes ved

$$\hat{S} = \sum (Y_i - \hat{\alpha}_1 - \hat{\beta} t_i)^2.$$

Videre introduserer jeg $Z_i = (Y_i - \alpha_1 - \beta t_i)/\sigma$ for $i = 1, 2, \ldots, n$, slik at $Z_1, Z_2, \ldots, Z_n$ er uavhengige og standardnormalfordelte variable. Responsvariabelen, Y_i, kan da uttrykkes ved $Y_i = \alpha_1 + \beta t_i + \sigma Z_i$. Dette kan brukes til å omskrive uttrykket for $\hat{\alpha}_1$. Siden $\bar{t} = 0$, er

$$\hat{\alpha}_1 = \bar{Y} = \alpha_1 + \sigma\bar{Z}.$$

Videre er

$$\begin{aligned}\hat{\beta} &= \frac{\sum Y_i t_i}{\sum t_i^2} = \frac{\sum \alpha_1 t_i + \beta t_i^2 + \sigma Z_i t_i}{\sum t_i^2} = \frac{\beta \sum t_i^2 + \sigma \sum Z_i t_i}{\sum t_i^2} \\ &= \beta + \frac{\sigma \sum Z_i t_i}{\sum t_i^2} = \beta + \frac{\sigma \sum Z_i t_i}{M},\end{aligned}$$

hvor $M = \sum t_i^2$.

Nå introduseres en passende ortogonal transformasjon av $Z_1, Z_2, \ldots, Z_n$. Jeg definerer

$$U_1 = \frac{1}{\sqrt{n}} Z_1 + \frac{1}{\sqrt{n}} Z_2 + \ldots + \frac{1}{\sqrt{n}} Z_n$$

og

$$U_2 = \frac{t_1}{\sqrt{M}} Z_1 + \frac{t_1}{\sqrt{M}} Z_2 + \ldots + \frac{t_1}{\sqrt{M}} Z_n.$$

Koeffisientene utgjør radvektorer med lengde 1 fordi $\sum (1/\sqrt{n})^2 = 1$ og $\sum \left(t_i/\sqrt{M}\right)^2 = M/M = 1$. Vektorene er ortogonale siden

$$\sum \frac{1}{\sqrt{n}} \frac{t_i}{\sqrt{M}} = \frac{1}{\sqrt{nM}} \sum t_i = 0.$$

Vi utvider disse vektorene til en komplett ortogonal matrise:

$$\mathbf{A} = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \frac{t_1}{\sqrt{M}} & \frac{t_2}{\sqrt{M}} & \cdots & \frac{t_n}{\sqrt{M}} \\ a_{31} & a_{32} & \cdots & a_{3n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}.$$

Vi definerer $U_i = \sum_{j=1}^{n} a_{ij} Z_j$ for $i = 3, 4, \ldots, n$. Fordi $\mathbf{A}$ er ortogonal, er $U_1, U_2, \ldots, U_n$ uavhengige og standardnormalfordelte. Det følger at

$$U_1 = \sqrt{n}\bar{Z} = \frac{\hat{\alpha}_1 - \alpha_1}{\sigma}\sqrt{n}$$

og

$$U_2 = \frac{1}{\sqrt{M}} \sum t_i Z_i = \frac{\hat{\beta} - \beta}{\sigma}\sqrt{M}.$$

Videre er

$$\begin{aligned} S = \sum (Y_i - \alpha_1 - \beta t_i)^2 &= \sum \left(Y_i - \hat{\alpha}_1 - \hat{\beta} t_i + (\hat{\alpha}_1 - \alpha_1) + (\hat{\beta} - \beta) t_i\right)^2 \\ &= \sum (Y_i - \hat{\alpha}_1 - \hat{\beta} t_i)^2 + \sum (\hat{\alpha}_1 - \alpha_1)^2 + \sum (\hat{\beta} - \beta)^2 t_i^2 \\ &\quad + 2(\hat{\alpha}_1 - \alpha_1) \sum (Y_i - \hat{\alpha}_1 - \hat{\beta} t_i) + 2(\hat{\beta} - \beta) \sum (Y_i - \hat{\alpha}_1 - \hat{\beta} t_i) t_i \\ &\quad + 2(\hat{\alpha}_1 - \alpha_1)(\hat{\beta} - \beta) \sum t_i \\ &= \hat{S} + n(\hat{\alpha}_1 - \alpha_1)^2 + M(\hat{\beta} - \beta)^2. \end{aligned}$$

Kryssleddene faller bort grunnet normallikningene og fordi $\sum t_i = 0$.

Vi har at

$$\begin{aligned} \frac{\hat{S}}{\sigma^2} &= \frac{S}{\sigma^2} - \frac{n(\hat{\alpha}_1 - \alpha_1)^2}{\sigma^2} - \frac{M(\hat{\beta} - \beta)^2}{\sigma^2} \\ &= \sum \left(\frac{Y_i - \alpha_1 - \beta t_i}{\sigma}\right)^2 - \left(\frac{(\hat{\alpha}_1 - \alpha_1)\sqrt{n}}{\sigma}\right)^2 - \left(\frac{\sqrt{M}(\hat{\beta} - \beta)}{\sigma}\right)^2 \\ &= \sum Z_i^2 - U_1^2 - U_2^2 = \sum U_i^2 - U_1^2 - U_2^2 = \sum_{i=3}^{n} U_i^2. \end{aligned}$$

Normalfordelingen til $\hat{\alpha}_1$ følger fra normalfordelingen til U_1. Vi har at $\mathrm{E}(\hat{\alpha}_1) = \alpha_1$ og $\mathrm{Var}(\hat{\alpha}_1) = \sigma^2/n$. Fra normalfordelingen til U_2 følger det at $\hat{\beta}$ er normalfordelt med

$$\mathrm{E}(\hat{\beta}) = \beta$$

og

$$\mathrm{Var}(\hat{\beta}) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}.$$

Uttrykket $\hat{S}/\sigma^2$ har en kjikvadratfordeling siden $\hat{S}/\sigma^2$ er en sum av $n - 2$ kvadrater av uavhengige normalfordelte variabler. Videre er $\hat{\alpha}_1, \hat{\beta}$ og $\hat{S}/\sigma^2$ uavhengige fordi $\hat{\alpha}_1$ uttrykkes kun ved hjelp av U_1, $\hat{\beta}$ uttrykkes bare ved hjelp av U_2 og $\hat{S}/\sigma^2$ uttrykkes kun ved hjelp av $U_3, U_4, \ldots, U_n$. Siden $\hat{\alpha} = \hat{\alpha}_1 - \hat{\beta}\bar{x}$, vil også $\hat{\alpha}$ være uavhengig av $\hat{S}/\sigma^2$, men ikke av $\hat{\beta}$.

Observatoren $\hat{\alpha}$ er en lineær kombinasjon av uavhengige, normalfordelte variable. Det følger at $\hat{\alpha}$ er normalfordelt. Forventningen er

$$\mathrm{E}(\hat{\alpha}) = \mathrm{E}(\hat{\alpha}_1 - \hat{\beta}\bar{x}) = \alpha_1 - \beta\bar{x} = \alpha.$$

Variansen er gitt ved

$$\mathrm{Var}(\hat{\alpha}) = \mathrm{Var}(\hat{\alpha}_1) + \bar{x}^2\mathrm{Var}(\hat{\beta}) = \frac{\sigma^2}{n} + \frac{\sigma^2\bar{x}^2}{\sum(x_i - \bar{x})^2} = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right).$$

Et kjent resultat sier at $T = Z/\sqrt{X^2/n}$ er t-fordelt med n frihetsgrader dersom Z er standardnormalfordelt, X^2 er kjikvadratfordelt med n frihetsgrader og Z og X er uavhengige. Inferens om α kan baseres på

$$T_1 = \frac{\dfrac{\hat{\alpha} - \alpha}{\sqrt{\sigma^2\left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right)}}}{\sqrt{\dfrac{\hat{S}}{\sigma^2(n-2)}}} = \frac{(\hat{\alpha} - \alpha)\sqrt{n-2}}{\sqrt{\hat{S}\left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right)}} \sim t(n-2).$$

Inferens om β kan utføres ved hjelp av

$$T_2 = \frac{\dfrac{\hat{\beta} - \beta}{\sqrt{\dfrac{\sigma^2}{\sum(x_i - \bar{x})^2}}}}{\sqrt{\dfrac{\hat{S}}{\sigma^2(n-2)}}} = \frac{(\hat{\beta} - \beta)\sqrt{n-2}}{\sqrt{\dfrac{\hat{S}}{\sum(x_i - \bar{x})^2}}} \sim t(n-2).$$

Til nå har forklaringsvariablene vært kjente konstanter. Av og til vil $x_1, x_2, \ldots, x_n$ være observerte verdier av stokastiske variable $X_1, X_2, \ldots, X_n$. En har behov for modeller hvor både den avhengige og den uavhengige variabelen er stokastiske. I vanlig lineær regresjon vil en fortsatt ofte tenke på x, den observerte verdien av X, som den uavhengige variabelen, og y, den observerte verdien av Y, som den avhengige variabelen. Altså ønsker en å predikere verdien til y etter å ha observert verdien til x. Det er da nærliggende å se på den betingede fordelingen til Y gitt $X = x$ [10].

Dersom en utfører lineær regresjon ved å bruke den betingede fordelingen til $Y_1, Y_2, \ldots, Y_n$ gitt $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, er vi i den samme situasjonen som tidligere, hvor $x_1, x_2, \ldots, x_n$ er faste størrelser. Estimatorene for α og β er de samme. Dersom forutsetningene for å utføre inferens er oppfylt, vil også testobservatorene være uforandret. I det videre arbeidet vil jeg nytte notasjonen $\mathrm{E}(Y_i) = \alpha + \beta x_i$, selv om notasjonen $\mathrm{E}(Y_i|X_i = x_i) = \alpha + \beta x_i$ i noen tilfeller vil være mer nøyaktig. Det spiller ingen rolle om $x_1, x_2, \ldots, x_n$ er observerte verdier av stokastiske variable så lenge vi betinger på disse [10].

9.1.3 Konsekvenser ved brudd på forutsetninger

Jeg skal vurdere regresjonsbaserte testmetoder for å identifisere publikasjonsbias i metaanalyser. Erfaringer fra rangkorrelasjonstestene viser at en bør vise forsiktighet ved utføring av tester dersom forutsetningene ikke er oppfylt. Jeg vil undersøke konsekvenser ved brudd på forutsetninger når en utfører inferens basert på en enkel, lineær regresjonsmodell. Noen forutsetninger vil diskuteres mer inngående enn andre. Dette gjelder særlig problemer knyttet til målefeil i forklaringsvariabelen. Denne diskusjonen vil være viktig på et senere stadium når jeg introduserer de aktuelle testene.

Ikke-lineær modell Hvis sammenhengen mellom Y og x ikke er lineær, kan man selvsagt ikke stole på resultatene en får ved å forsøke å tilpasse en lineær modell til dataene. Det er vanskelig å si noe generelt om konsekvensene, de vil avhenge av hver enkelt situasjon. I noen tilfeller kan passende transformasjoner omforme en ikke-lineær sammenheng til en sammenheng som er lineær, eller i alle fall tilnærmet lineær. Jeg vil ikke utdype denne problemstillingen. Den interesserte leser henvises til Draper og Smith [12], hvor forfatterne gir en grei og ikke for omfattende innføring i emnet.

Målefeil i forklaringsvariabelen Den uavhengige variabelen skal være målt uten feil. Dersom forklaringsvariabelen inneholder målefeil, kan dette medføre skjevhet i estimatet for stigningstallet. Utstrekningen og retningen på denne biasen vil avhenge av variansen og kovariansen til den virkelige verdien av den uavhengige variabelen og målefeilen. Dersom ingen korrelasjon eksisterer mellom målefeilen og den virkelige verdien av den uavhengige variabelen, vil en få en fortynning av stigningstallet [49]. Dette kalles attenuation bias eller fortynningsskjevhet. Skjevhet i estimatet for stigningstallet kan medføre skjevhet i estimatet for skjæringspunktet.

Jeg vil vise dette under visse forutsetninger. Vi lar $Y_1, Y_2, \ldots, Y_n$ være uavhengige og normalfordelte variable. Videre er $Y_i = \alpha + \beta X_i + \epsilon_i$ for $i = 1, 2, \ldots, n$, hvor X_i er den sanne verdien av forklaringsvariabelen. Denne verdien er ukjent. Vi kjenner kun $X_i' = X_i + e_i$, hvor e_i er målefeilen til X_i. Videre antar vi at X_i' er en forventningsrett estimator for den uobserverte stokastiske variabelen X_i, slik at e_i er en stokastisk variabel med forventning lik null. I første omgang antar vi også at ϵ_i, X_i og e_i er gjensidig uavhengige og normalfordelte for alle i. I denne utledningen er X_i en stokastisk variabel. Tilsvarende resultater om fortynningsskjevhet kan utledes om X_i er en fast størrelse [10].

Jeg starter med å definere α og β. Dersom forutsetningene er oppfylt, er X og Y som kjent bivariat normalfordelt. Det følger at også den betingede fordelingen til Y gitt $X = x$ er

normalfordelt. Fra den betingede bivariate normalfordelingen vet vi at

$$\mathrm{E}(Y|X=x) = \mathrm{E}(Y) - \frac{\mathrm{Cov}(X,Y)\mathrm{E}(X)}{\mathrm{Var}(X)} + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}x.$$

Regresjonen fra Y på $X = x$ er lineær. En naturlig definisjon av parametrene α og β er gitt ved

$$\alpha = \mathrm{E}(Y) - \frac{\mathrm{Cov}(X,Y)\mathrm{E}(X)}{\mathrm{Var}(X)}$$

og

$$\beta = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}.$$

I tillegg vil variansen til Y gitt $X = x$ være uavhengig av x fordi

$$\mathrm{Var}(Y|X=x) = \mathrm{Var}(Y)\left(1 - \frac{\mathrm{Cov}^2(X,Y)}{\mathrm{Var}(X)\mathrm{Var}(Y)}\right).$$

Jeg vil senere ha behov for å vite den bivariate fordelingen til X_i' og Y_i og utleder derfor denne. Jeg definerer

$$\boldsymbol{\delta}_i = \begin{pmatrix} \epsilon_i \\ X_i \\ e_i \end{pmatrix}.$$

Det følger at $\boldsymbol{\delta}_i$ er multivariat normalfordelt. Forventningen til $\boldsymbol{\delta}_i$, $\boldsymbol{\mu}$, er

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ \mathrm{E}(X) \\ 0 \end{pmatrix},$$

mens kovariansmatrisen, $\boldsymbol{\Sigma}$, kan uttrykkes ved

$$\boldsymbol{\Sigma} = \begin{pmatrix} \mathrm{Var}(\epsilon) & 0 & 0 \\ 0 & \mathrm{Var}(X) & 0 \\ 0 & 0 & \mathrm{Var}(e) \end{pmatrix}.$$

Videre defineres

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & \beta & 0 \end{pmatrix}$$

og

$$\boldsymbol{\alpha} = \begin{pmatrix} 0 \\ \alpha \end{pmatrix},$$

slik at

$$\begin{pmatrix} X_i' \\ Y_i \end{pmatrix} = \boldsymbol{\alpha} + \mathbf{A}\boldsymbol{\delta}_i.$$

Resultater fra multivariat analyse gir at

$$\begin{pmatrix} X_i' \\ Y_i \end{pmatrix} \sim N_2(\boldsymbol{\alpha} + \mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T), \qquad i = 1, 2, \ldots, n,$$

hvor $\mathbf{A}^T$ er den transponerte til $\mathbf{A}$. Her er

$$\boldsymbol{\alpha} + \mathbf{A}\boldsymbol{\mu} = \begin{pmatrix} \mathrm{E}(X) \\ \alpha + \beta \mathrm{E}(X) \end{pmatrix}$$

og

$$\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T = \begin{pmatrix} \mathrm{Var}(X) + \mathrm{Var}(e) & \beta \mathrm{Var}(X) \\ \beta \mathrm{Var}(X) & \beta^2 \mathrm{Var}(X) + \mathrm{Var}(\epsilon) \end{pmatrix}.$$

La α' og β' betegne regresjonskoeffisientene i regresjonen fra Y på $X' = x'$. Jeg estimerer α' og β' ved minste kvadraters metode og minimerer $S = \sum(y_i - \alpha' - \beta' x_i')^2$. Det følger at

$$\hat{\alpha}' = \bar{Y} - \hat{\beta}' \bar{X}'$$

og

$$\hat{\beta}' = \frac{\sum Y_i (X_i' - \bar{X}')}{\sum (X_i' - \bar{X}')^2}.$$

For å vise fortynningsskjevhet og skjevhet i estimatet for skjæringspunktet, vil jeg vise at $\hat{\alpha}'$ og $\hat{\beta}'$ ikke er konsistente estimatorer for henholdsvis α og β. En konsistent estimator kan enkelt forklares ved at jo større sampelstørrelse en har, desto mer nøyaktige og presise vil estimatene være.

Siden de n parene $(Y_1, X_1'), (Y_2, X_2'), \ldots, (Y_n, X_n')$ er uavhengige og hvert par har den samme bivariate normalfordelingen, vil $\hat{\beta}'$ konvergere i sannsynlighet mot $\mathrm{Cov}(Y, X')/\mathrm{Var}(X')$. Dette følger fra Khinchins setning sammen med Slutskys setning, som definert hos Meen og Heuch [37].

Jeg erstatter X' med $X + e$. Siden vi antar at e er uavhengig av både X og Y, får vi at

$$\begin{aligned} \operatorname*{plim}_{n\to\infty} \hat{\beta}' &= \frac{\mathrm{Cov}(Y, X')}{\mathrm{Var}(X')} = \frac{\mathrm{Cov}(Y, X + e)}{\mathrm{Var}(X + e)} = \frac{\mathrm{Cov}(Y, X) + \mathrm{Cov}(Y, e)}{\mathrm{Var}(X + e)} \\ &= \frac{\mathrm{Cov}(Y, X)}{\mathrm{Var}(X) + \mathrm{Var}(e)} = \frac{\beta}{1 + \dfrac{\mathrm{Var}(e)}{\mathrm{Var}(X)}} = \frac{\beta}{1 + \lambda}, \end{aligned}$$

hvor $\lambda = \mathrm{Var}(e)/\mathrm{Var}(X)$. Se Casella og Berger [10] for liknende beregninger. En ser at målefeil i den uavhengige variabelen fører til at $\hat{\beta}'$ er en inkonsistent estimator for β. Videre har $\operatorname*{plim}_{n\to\infty} \hat{\beta}'$ samme fortegn som β, men er biased mot null siden $\beta/(1+\lambda) < \beta$. Dette kalles fortynningsbias.

Vi husker videre at $\hat{\alpha}' = \bar{Y} - \bar{X}'\hat{\beta}'$. Vi vet at $\bar{Y}$ konvergerer i sannsynlighet mot $\mathrm{E}(Y) = \alpha + \beta\mathrm{E}(X')$, og at $\bar{X}'$ konvergerer i sannsynlighet mot $\mathrm{E}(X')$. Dersom $\hat{\beta}'$ konvergerer i sannsynlighet mot $\beta/(1+\lambda)$, vil

$$\hat{\alpha}' \xrightarrow{P} \alpha + \beta\mathrm{E}(X') - \mathrm{E}(X')\frac{\beta}{1+\lambda} = \alpha + \beta\mathrm{E}(X')\left(\frac{\lambda}{1+\lambda}\right).$$

Generelt er $\hat{\alpha}'$ derfor en inkonsistent estimator for α. Vi får negativ skjevhet dersom β og $\mathrm{E}(X')$ har motsatt fortegn. Biasen er positiv dersom β og $\mathrm{E}(X')$ har samme fortegn og forsvinner om β eller $\mathrm{E}(X')$ er lik null.

Det er også et poeng å vise at $\hat{\alpha}'$ og $\hat{\beta}'$ ikke er forventningsrette estimatorer for henholdsvis α og β. Innledningsvis i denne seksjonen viste jeg at Y og X' er bivariat normalfordelt. Det følger at også den betingede fordelingen til Y gitt $X' = x'$ er normalfordelt. En naturlig definisjon av parametrene $\hat{\alpha}'$ og $\hat{\beta}'$ er gitt ved henholdsvis

$$\alpha' = \mathrm{E}(Y) - \frac{\mathrm{Cov}(X', Y)\mathrm{E}(X')}{\mathrm{Var}(X')}$$

og

$$\beta' = \frac{\mathrm{Cov}(X', Y)}{\mathrm{Var}(X')}.$$

Kravet om at X skal være normalfordelt er viktig for å sikre at regresjonen fra Y på $X' = x'$ er lineær. Regelen om dobbelforventning, samt teori fra regresjonsanalyse og den multivariate normalfordelingen, gir at

$$\mathrm{E}(\hat{\beta}') = \mathrm{EE}(\hat{\beta}'|X_1', \ldots, X_n') = \mathrm{E}(\beta') = \beta' = \frac{\mathrm{Cov}(X', Y)}{\mathrm{Var}(X')} = \frac{\beta}{1+\lambda}.$$

På samme måte får jeg at

$$\begin{aligned}\mathrm{E}(\hat{\alpha}') &= \mathrm{EE}(\hat{\alpha}'|X_1', \ldots, X_n') = \mathrm{E}(\alpha') = \alpha' = \mathrm{E}(Y) - \frac{\mathrm{Cov}(X', Y)\mathrm{E}(X')}{\mathrm{Var}(X')} \\ &= \alpha + \beta\mathrm{E}(X')\left(\frac{\lambda}{1+\lambda}\right).\end{aligned}$$

Jeg beveger meg over i en annen situasjon. Forutsetningene gitt innledningsvis i denne seksjonen gjelder fortsatt, bortsett fra at jeg ikke lenger krever at X og e skal være uavhengige. Jeg antar at $(X_1', Y_1), (X_2', Y_2), \ldots, (X_n', Y_n)$ er n uavhengige par som alle har den samme bivariate fordelingen. Nå konvergerer $\hat{\beta}'$ i sannsynlighet mot

$$\begin{aligned}\operatorname*{plim}_{n\to\infty}\hat{\beta}' &= \frac{\mathrm{Cov}(Y, X')}{\mathrm{Var}(X')} = \frac{\mathrm{Cov}(Y, X+e)}{\mathrm{Var}(X+e)} = \frac{\mathrm{Cov}(Y, X) + \mathrm{Cov}(Y, e)}{\mathrm{Var}(X) + \mathrm{Var}(e) + 2\mathrm{Cov}(X, e)} \\ &= \frac{\mathrm{Cov}(Y, X) + \mathrm{Cov}(\alpha + \beta X + \epsilon\ ,\ e)}{\mathrm{Var}(X) + \mathrm{Var}(e) + 2\mathrm{Cov}(X, e)} = \frac{\beta\mathrm{Var}(X) + \beta\mathrm{Cov}(X, e)}{\mathrm{Var}(X) + \mathrm{Var}(e) + 2\mathrm{Cov}(X, e)} \\ &= \frac{\beta\big(\mathrm{Var}(X) + \mathrm{Cov}(X, e)\big)}{\mathrm{Var}(X) + \mathrm{Var}(e) + 2\mathrm{Cov}(X, e)}.\end{aligned}$$

Det følger at

$$\hat{\alpha}' \xrightarrow{P} \alpha + \beta \mathrm{E}(X') \left(1 - \frac{\mathrm{Var}(X) + \mathrm{Cov}(X, e)}{\mathrm{Var}(X) + \mathrm{Var}(e) + 2\mathrm{Cov}(X, e)}\right).$$

Videre er

$$\mathrm{E}(\hat{\beta}') = \frac{\beta\big(\mathrm{Var}(X) + \mathrm{Cov}(X, e)\big)}{\mathrm{Var}(X) + \mathrm{Var}(e) + 2\mathrm{Cov}(X, e)}$$

og

$$\mathrm{E}(\hat{\alpha}') = \alpha + \beta \mathrm{E}(X') \left(1 - \frac{\mathrm{Var}(X) + \mathrm{Cov}(X, e)}{\mathrm{Var}(X) + \mathrm{Var}(e) + 2\mathrm{Cov}(X, e)}\right).$$

Utledningen følger ved tilsvarende beregninger som for situasjonen hvor X og e er uavhengige. Hverken $\hat{\alpha}'$ eller $\hat{\beta}'$ er forventningsrette estimatorer for henholdsvis α og β. Beregningene bekrefter at vi får en skjevhet i estimatet for stigningstallet. Utstrekningen og retningen på denne skjevheten avhenger av variansen og kovariansen til den virkelige verdien av forklaringsvariabelen og målefeilen. Dette medfører bias i estimatet for skjæringspunktet.

Resultatene understøtter at testobservatorene generelt må utbedres hvis en ønsker å utføre inferens om α og β dersom forklaringsvariabelen er målt med feil. Bruken av t-fordeling vil generelt ikke gi korrekt nivå for testing av $H_0 : \alpha = \alpha_0$ eller $H_0 : \beta = \beta_0$. Unntak finnes. Hypotesetesting om $\beta = 0$ vil ikke inneholde bias dersom vi har uavhengige målefeil. Dette følger fordi biasen i $\hat{\beta}'$ er en faktor multiplisert med β. Målefeil vil likevel redusere styrken til denne testen [18].

Biasen i $\hat{\beta}'$ er en funksjon av $\lambda = \mathrm{Var}(e)/\mathrm{Var}(X) = \mathrm{Var}(e)/\big(\mathrm{Var}(X') - \mathrm{Var}(e)\big)$ dersom X og e er uavhengige. La S_e^2, $S_{X'}^2$ være estimatorer for henholdsvis $\mathrm{Var}(e)$ og $\mathrm{Var}(X')$. Vi kan estimere λ ved $\hat{\lambda} = S_e^2/(S_{X'}^2 - S_e^2)$. Fra dette estimatet kan en vurdere om biasen er neglisjerbar eller ikke. Dersom biasen ikke er neglisjerbar, kan en estimere β ved $\hat{\beta}'(1 + \hat{\lambda})$. Dette vil minske skjevheten [49]. Frost og Thompson [17] diskuterer ulike korrigeringsmetoder.

Jeg inkluderer en liten digresjon til slutt. Vi vet at den avhengige variabelen, Y_i, er utsatt for feilen ϵ_i, hvor $\mathrm{E}(\epsilon_i) = 0$ og $\mathrm{Var}(\epsilon_i) = \sigma^2$. Den uavhengige variabelen, x_i, antas nå å være målt uten feil. Hvis ϵ_i ikke tar hensyn til usikkerhet knyttet til målefeil, må vi undersøke konsekvensene av denne tilleggsfeilen. Vi lar Y_i betegne den sanne, ukjente verdien av den avhengige variabelen. Vi observerer Y_i' og har sammenhengen $Y_i' = Y_i + e_i$, hvor vi antar at e_i er normalfordelt med forventning lik null. Innsetting i det vanlige, lineære uttrykket $Y_i = \alpha + \beta x_i + \epsilon_i$ gir $Y_i' = \alpha + \beta x_i + \epsilon_i + e_i$. Vi ser at målefeil i Y_i kun forandrer tolkningen av feilleddet, da $\mathrm{Var}(Y_i') = \mathrm{Var}(\epsilon_i) + \mathrm{Var}(e_i)$ dersom ϵ_i og e_i er uavhengige. Testobservatorene må ta hensyn til denne ekstra variasjonen. Forventningen til Y_i' vil være lik forventningen til Y_i.

Homoskedastisitet Det er videre et krav om konstant varians, kalt homoskedastisitet. Dersom ikke dette er oppfylt, vil estimatorene for α og β likevel forbli de samme. Dette ser en lett fra utledningen av minste kvadraters estimatorer for α og β, gitt innledningvis i kapittelet. Begge estimatorene vil være forventningsrette for henholdsvis α og β. Variansen til estimatorene vil derimot endres, slik at vi må utbedre testobservatorene.

La $Y_1, Y_2, \ldots, Y_n$ være uavhengige. Vi kan ikke lenger anta at $\mathrm{Var}(Y_i) = \mathrm{Var}(Y_j)$ for $i \neq j$. Dette medfører at

$$\mathrm{Var}(\hat{\beta}) = \sum \frac{(x_i - \bar{x})^2 \mathrm{Var}(Y_i)}{\left(\sum (x_i - \bar{x})^2\right)^2}.$$

Det følger at

$$\begin{aligned}\mathrm{Var}(\hat{\alpha}) &= \frac{1}{n^2}\sum \mathrm{Var}(Y_i) + \bar{x}^2 \mathrm{Var}(\hat{\beta}) - \frac{2}{n}\bar{x}\mathrm{Cov}\left(\sum Y_i, \hat{\beta}\right) \\ &= \frac{1}{n^2}\sum \mathrm{Var}(Y_i) + \bar{x}^2 \frac{\sum (x_i - \bar{x})^2 \mathrm{Var}(Y_i)}{\left(\sum (x_i - \bar{x})^2\right)^2} - \frac{2\bar{x}}{n\sum (x_i - \bar{x})^2}\sum (x_i - \bar{x})\mathrm{Var}(Y_i).\end{aligned}$$

Vektet regresjon tillater heteroskedastisitet. Jeg vil se på denne situasjonen siden.

Ikke-uavhengige feilledd Hva om feilleddene ikke er uavhengige? Mye kan gå galt i utledningen av testobservatorene. Eksempelvis er variansen til $\hat{\beta}$ gitt ved

$$\mathrm{Var}(\hat{\beta}) = \mathrm{Var}\left(\frac{\sum (x_i - \bar{x})Y_i}{\sum (x_i - \bar{x})^2}\right) = \frac{1}{\left(\sum (x_i - \bar{x})^2\right)^2}\mathrm{Var}\left(\sum (x_i - \bar{x})Y_i\right).$$

Fordi Y_i ikke lenger kan antas å være uavhengig av Y_j, vil en ikke kunne anta at $\mathrm{Var}(\hat{\beta}) = \sum (x_i - \bar{x})^2 \mathrm{Var}(Y_i) / \left(\sum (x_i - \bar{x})^2\right)^2$. Dette vil selvsagt få konsekvenser for størrelsen på testobservatoren i utvalget for både α og β.

Normalfordelte feilledd Forutsetningen om normalfordelte feilledd gjelder spesielt for hypotesetestingen av de estimerte størrelsene i regresjonsmodellen, og angår således ikke om våre estimerte størrelser er korrekte. I små utvalg trenger vi normalfordelte feilledd for at feilmarginer og signifikansnivåer skal bli korrekte. Sentralgrenseteoremet sikrer at problemet minimeres om denne forutsetningen brytes i store utvalg [51].

Underspesifisert regresjonsmodell Vi antar modellen $\mathrm{E}(Y_i) = \alpha + \beta x_i$. Alle faktorer som påvirker Y_i, bortsett fra x_i, er samlet i restleddet, ϵ_i. Minste kvadraters metode gir som vanlig

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$$

og

$$\hat{\beta} = \frac{\sum Y_i (x_i - \bar{x})}{\sum (x_i - \bar{x})^2}.$$

Estimatorene er forventningsrette for α og β.

Hva om modellen ikke er korrekt spesifisert? Vi tenker vi har utelatt relevante årsaker fra modellen, slik at den sanne, underliggende sammenhengen er $\mathrm{E}(Y_i) = \alpha + \beta x_i + \eta z_i$. Modellen, $\mathrm{E}(Y_i) = \alpha + \beta x_i$, er da underspesifisert. Dette medfører at

$$\begin{aligned}\mathrm{E}(\hat{\beta}) &= \frac{\sum(x_i - \bar{x})\mathrm{E}(Y_i)}{\sum(x_i - \bar{x})^2} = \frac{\sum(x_i - \bar{x})(\alpha + \beta x_i + \eta z_i)}{\sum(x_i - \bar{x})^2} \\ &= \beta + \eta\frac{\sum z_i(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}\end{aligned}$$

og

$$\begin{aligned}\mathrm{E}(\hat{\alpha}) = \mathrm{E}(\bar{Y} - \hat{\beta}\bar{x}) &= \frac{1}{n}\sum(\alpha + \beta x_i + \eta z_i) - \bar{x}\left(\beta + \eta\frac{\sum z_i(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}\right) \\ &= \alpha + \eta\bar{z} - \bar{x}\eta\frac{\sum z_i(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}.\end{aligned}$$

Estimatorene er ikke lenger forventningsrette for α og β. Heuristisk antar vi at $\hat{\beta} \xrightarrow{P} \beta + \eta\mathrm{Cov}(x, z)/\mathrm{Var}(x)$. Biasen i estimatet for β forsvinner dersom x_i og z_i er ukorrelerte [47], men vil fortsatt være et problem for $\hat{\alpha}$.

Seber [47] viser til mulige utfordringer dersom modellen er overspesifisert. Dette er et tema litt på siden av saken, og jeg vil ikke behandle disse problemene her. I samme bok argumenterer forfatteren for at problemet med ukorrekt spesifisering av modellen kun vil forekomme om forklaringsvariablene er faste størrelser og ikke observerte verdier av stokastiske variable. Vi har modellen

$$\begin{aligned}Y &= \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_s \\ &= \beta_0 + \beta_1 X_1 + \ldots + \beta_r X_r + \delta, \quad r < s,\end{aligned}$$

hvor X_j er stokastiske variable, målt uten feil, og $\mathrm{E}(X_j) = \theta_j$ for $j = 1, 2, \ldots, s$. Alle påvirkninger på Y bortsett fra $X_1, X_2, \ldots, X_r$ er samlet i restleddet δ. Samlingen av $X_1, X_2, \ldots, X_r$ er uavhengig av samlingen $X_{r+1}, X_{r+2}, \ldots, X_s$

For den $i$te uavhengige repetisjonen av et eksperiment har vi at

$$\begin{aligned}Y_i &= \beta_0 + \beta_1 X_{i1} + \ldots + \beta_s X_{is} \\ &= (\beta_0 + \beta_1\bar{X}_{.1} + \ldots + \beta_r\bar{X}_{.r} + \beta_{r+1}\theta_{r+1} + \ldots + \beta_s\theta_s) \\ &\quad + \beta_1(X_{i1} - \bar{X}_{.1}) + \ldots + \beta_r(X_{ir} - \bar{X}_{.r}) \\ &\quad + \big(\beta_{r+1}(X_{i,r+1} - \theta_{r+1}) + \ldots + \beta_s(X_{is} - \theta_s)\big) \\ &= \alpha + \beta_1(X_{i1} - \bar{X}_{.1}) + \ldots + \beta_r(X_{ir} - \bar{X}_{.r}) + \epsilon, \qquad i = 1, 2, \ldots, n.\end{aligned}$$

Her betegner $\bar{X}_{.j} = \sum_i X_{ij}$. Videre er α en stokastisk variabel. Fordi $\mathrm{E}(X_{ij}) = \theta_j$, er $\mathrm{E}(\epsilon) = 0$.

Vi ser nå på den betingede forventningen til Y_i gitt $X_{ij} = x_{ij}$ for $i = 1, 2, \ldots, n$ og $j = 1, 2, \ldots, r$. Dette gir

$$\mathrm{E}(Y_i|\mathbf{X} = \mathbf{x}) = \alpha + \beta_1(x_{i1} - \bar{x}_{.1}) + \ldots + \beta_r(x_{ir} - \bar{x}_{.r}),$$

hvor $\mathbf{X}$ består av $X_{i1}, X_{i2}, \ldots, X_{ir}$ for $i = 1, 2, \ldots, n$. Betinget vil α være konstant. Vi leter etter en modell som reduserer ϵ til et rimelig nivå. Siden r er vilkårlig, har vi at $\mathrm{E}(\epsilon) = 0$ uansett hvor mange forklaringsvariabler vi inkluderer i modellen.

9.2 Vektet lineær regresjon

Vektet lineær regresjon er for mange mest kjent som en regresjonsmetode som tar hensyn til at variansen til Y_i ikke nødvendigvis er lik variansen til Y_j, for $i \neq j$. Usikre observasjoner, det vil si observasjoner med store standardavvik, bør vektlegges i mindre grad enn troverdige observasjoner med lite standardavvik. Dette kan oppnås dersom vi vekter hver responsvariabel med inversen til denne variabelens varians.

Også i denne modellen skal $Y_1, Y_2, \ldots, Y_n$ være uavhengige og normalfordelte variable. Vi har sammenhengen $\mathrm{E}(Y_i) = \alpha + \beta x_i$. Nå er $\mathrm{Var}(Y_i) = \sigma^2/w_i$ for $i = 1, 2, \ldots, n$, hvor w_i er et kjent, positivt tall. Både w_i og $x_1, x_2, \ldots, x_n$ skal være målt uten feil.

9.2.1 Vektet minste kvadraters metode og utledning av testobservatorer

Utledningen av punktestimatorene for α og β ved hjelp av vektet minste kvadraters metode følger samme framgangsmåte som utledningen ved minste kvadraters metode. Definisjoner og beregninger blir likevel noe annerledes om en også ønsker å utføre inferens. Jeg velger å gi beviset i sin helhet. Jeg har ikke funnet litteratur som utleder testobservatorene.

Jeg introduserer $t_i = w_i(x_i - \bar{x}_w)$, hvor $\bar{x}_w = \sum w_i x_i / \sum w_i$. Da blir

$$\begin{aligned}\sum t_i &= \sum w_i(x_i - \bar{x}_w) = \sum w_i x_i - \bar{x}_w \sum w_i \\ &= \sum w_i x_i - \left(\sum w_i\right)\left(\frac{\sum w_i x_i}{\sum w_i}\right) = \sum w_i x_i - \sum w_i x_i = 0.\end{aligned}$$

Dette gir $x_i = t_i w_i^{-1} + \bar{x}_w$. Forventningen til Y_i kan uttrykkes ved $\mathrm{E}(Y_i) = \alpha + \beta x_i = \alpha + \beta\bar{x}_w + \beta t_i w_i^{-1}$. Jeg definerer $\alpha_1 = \alpha + \beta\bar{x}_w$ slik at $\mathrm{E}(Y_i) = \alpha_1 + \beta t_i w_i^{-1}$.

Vektet minste kvadraters estimatorer kan finnes ved å minimere

$$S = \sum w_i(y_i - \alpha_1 - \beta t_i w_i^{-1})^2.$$

Jeg deriverer først S med hensyn på α_1, deretter med hensyn på β. Normallikningene er gitt ved

$$\frac{\partial S}{\partial \alpha_1} = (-2)\sum w_i(y_i - \alpha_1 - \beta t_i w_i^{-1}) = 0$$

og

$$\frac{\partial S}{\partial \beta} = (-2)\sum t_i(y_i - \alpha_1 - \beta t_i w_i^{-1}) = 0.$$

Løser vi normallikningene, får vi estimatorene

$$\hat{\alpha}_1 = \frac{\sum w_i Y_i}{\sum w_i} = \bar{Y}_w$$

og

$$\hat{\beta} = \frac{\sum t_i Y_i}{\sum t_i^2 w_i^{-1}}.$$

Jeg definerer $M = \sum t_i^2 w_i^{-1}$ slik at $\hat{\beta} = \sum t_i Y_i / M$. Minimert kvadratsum er definert ved

$$\hat{S} = \sum w_i (Y_i - \hat{\alpha}_1 - \hat{\beta} t_i w_i^{-1})^2.$$

Jeg ønsker uavhengige standardnormalfordelte variable og standardiserer Y_i ved

$$Z_i = \frac{Y_i - \alpha_1 - \beta t_i w_i^{-1}}{\sigma\sqrt{w_i^{-1}}}, \quad i = 1, 2, \ldots, n.$$

Jeg kan nå uttrykke Y_i ved

$$Y_i = Z_i \sigma \sqrt{w_i^{-1}} + \alpha_1 + \beta t_i w_i^{-1}.$$

Dette impliserer at

$$\begin{aligned}\hat{\alpha}_1 = \bar{Y}_w &= \frac{\sum w_i Z_i \sigma \sqrt{w_i^{-1}}}{\sum w_i} + \frac{\alpha_1 \sum w_i}{\sum w_i} + \frac{\beta \sum t_i w_i^{-1} w_i}{\sum w_i} \\ &= \alpha_1 + \sigma \frac{\sum w_i^{1/2} Z_i}{\sum w_i}\end{aligned}$$

og

$$\begin{aligned}\hat{\beta} = \frac{\sum t_i Y_i}{M} &= \frac{1}{M}\sum (t_i Z_i \sigma \sqrt{w_i^{-1}} + t_i \alpha_1 + \beta t_i^2 w_i^{-1}) \\ &= \beta + \frac{\sigma}{M}\sum t_i Z_i \sqrt{w_i^{-1}}.\end{aligned}$$

Videre introduserer jeg passende ortogonale transformasjoner av $Z_1, Z_2, \ldots, Z_n$. Det viser seg at

$$U_1 = \frac{Z_1}{\sqrt{w_1^{-1}\sum w_i}} + \frac{Z_2}{\sqrt{w_2^{-1}\sum w_i}} + \ldots + \frac{Z_n}{\sqrt{w_n^{-1}\sum w_i}}$$

og

$$U_2 = \frac{t_1 Z_1 \sqrt{w_1^{-1}}}{\sqrt{M}} + \frac{t_2 Z_2 \sqrt{w_2^{-1}}}{\sqrt{M}} + \ldots + \frac{t_n Z_n \sqrt{w_n^{-1}}}{\sqrt{M}}$$

er to ortonormale vektorer fordi

$$\sum \left(\frac{1}{\sqrt{w_i^{-1} \sum w_i}} \right)^2 = \sum \frac{1}{w_i^{-1} \sum w_i} = \frac{1}{\sum w_i} \sum w_i = 1,$$

$$\sum \left(\frac{t_i \sqrt{w_i^{-1}}}{\sqrt{M}} \right)^2 = \frac{1}{M} \sum t_i^2 w_i^{-1} = \frac{M}{M} = 1$$

og

$$\sum \frac{1}{\sqrt{w_i^{-1} \sum w_i}} \frac{t_i \sqrt{w_i^{-1}}}{\sqrt{M}} = \frac{1}{\sqrt{M \sum w_i}} \sum t_i = 0.$$

Jeg utvider disse vektorene til en komplett, ortogonal matrise:

$$\mathbf{A} = \begin{pmatrix} \frac{1}{\sqrt{w_1^{-1} \sum w_i}} & \frac{1}{\sqrt{w_2^{-1} \sum w_i}} & \cdots & \frac{1}{\sqrt{w_n^{-1} \sum w_i}} \\ \frac{t_1 \sqrt{w_1^{-1}}}{\sqrt{M}} & \frac{t_2 \sqrt{w_2^{-1}}}{\sqrt{M}} & \cdots & \frac{t_n \sqrt{w_n^{-1}}}{\sqrt{M}} \\ a_{31} & a_{32} & \cdots & a_{3n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}.$$

Videre defineres $U_3, U_4, \ldots, U_n$ ved den ortogonale transformasjonen $U_i = \sum_{j=1}^{n} a_{ij} Z_j$, $i = 3, 4, \ldots, n$. Det følger at $U_1, U_2, \ldots, U_n$ er uavhengige og standardnormalfordelte variable.

Det viser seg at $\hat{\alpha}_1$, $\hat{\beta}$ og $\hat{S}$ kan uttrykkes ved hjelp av de transformerte variablene. Vi har at

$$\frac{\hat{\alpha}_1 - \alpha_1}{\sigma} \sqrt{\sum w_i} = U_1$$

og

$$\frac{\sqrt{M}(\hat{\beta} - \beta)}{\sigma} = U_2.$$

Disse uttrykkene følger uten anstrengelser. Videre er $\hat{S}/\sigma^2 = \sum_{i=3}^{n} U_i^2$. Noen mellomregninger kan være på sin plass for å vise dette. Vi har at

$$\begin{aligned} S &= \sum w_i (Y_i - \alpha_1 - \beta t_i w_i^{-1})^2 = \sum w_i \big((Y_i - \hat{\alpha}_1 - \hat{\beta} t_i w_i^{-1}) + (\hat{\alpha}_1 - \alpha_1) + (\hat{\beta} - \beta) t_i w_i^{-1} \big)^2 \\ &= \sum w_i (Y_i - \hat{\alpha}_1 - \hat{\beta} t_i w_i^{-1})^2 + \sum w_i (\hat{\alpha}_1 - \alpha_1)^2 + \sum w_i (\hat{\beta} - \beta)^2 t_i^2 w_i^{-2} \\ &\quad + 2(\hat{\alpha}_1 - \alpha_1) \sum w_i (Y_i - \hat{\alpha}_1 - \hat{\beta} t_i w_i^{-1}) + 2(\hat{\beta} - \beta) \sum w_i t_i w_i^{-1} (Y_i - \hat{\alpha}_1 - \hat{\beta} t_i w_i^{-1}) \\ &\quad + 2(\hat{\alpha}_1 - \alpha_1)(\hat{\beta} - \beta) \sum t_i w_i^{-1} w_i. \end{aligned}$$

Kryssleddene faller bort fordi $\sum t_i = 0$ og på grunn av normallikningene. Det følger at

$$\hat{S} = S - (\hat{\alpha}_1 - \alpha_1)^2 \sum w_i - (\hat{\beta} - \beta)^2 M.$$

Bruker vi dette resultatet, ser vi at

$$\begin{aligned}\frac{\hat{S}}{\sigma^2} &= \frac{S}{\sigma^2} - \frac{(\hat{\alpha}_1 - \alpha_1)^2}{\sigma^2} \sum w_i - \frac{(\hat{\beta} - \beta)^2}{\sigma^2} M \\ &= \sum Z_i^2 - U_1^2 - U_2^2 = \sum U_i^2 - U_1^2 - U_2^2 = \sum_{i=3}^{n} U_i^2.\end{aligned}$$

Normalfordelingen til $\hat{\alpha}_1$ og $\hat{\beta}$ følger fra normalfordelingen til U_1 og U_2 henholdvis. Kjikvadratfordelingen til $\hat{S}/\sigma^2$ er en konsekvens av at variabelen er en sum av $n-2$ kvadrater av uavhengige, standardnormalfordelte variable. Uavhengigheten mellom $\hat{\alpha}_1$, $\hat{\beta}$ og $\hat{S}/\sigma^2$ følger fordi $\hat{\alpha}_1$ kan uttrykkes som en funksjon av U_1, $\hat{\beta}$ kan uttrykkes som en funksjon av U_2 og $\hat{S}/\sigma^2$ er en funksjon av $U_3, U_4, \ldots, U_n$.

Jeg er i denne oppgaven interessert i $\hat{\alpha}$ og ikke i $\hat{\alpha}_1$. En vektet minste kvadraters estimator for $\hat{\alpha}$ finner jeg ved å minimere $S = \sum w_i(y_i - \alpha - \beta x_i)^2$. Estimatoren er gitt ved $\hat{\alpha} = \bar{Y}_w - \hat{\beta}\bar{x}_w$, og kan uttrykkes ved hjelp av U_1 og U_2. Punktestimatoren er uavhengig av $\hat{S}/\sigma^2$. Videre er $\hat{\alpha}$ en lineær kombinasjon av uavhengige, normalfordelte variable og er selv normalfordelt. Forventningen til $\hat{\alpha}$ er gitt ved

$$\mathrm{E}(\hat{\alpha}) = \mathrm{E}(\hat{\alpha}_1) - \mathrm{E}(\hat{\beta}\bar{x}_w) = \alpha.$$

Fordi $\hat{\alpha}_1$ og $\hat{\beta}$ er uavhengige, er variansen til $\hat{\alpha}$ gitt ved

$$\mathrm{Var}(\hat{\alpha}) = \mathrm{Var}(\hat{\alpha}_1) + \bar{x}_w^2 \mathrm{Var}(\hat{\beta}) = \sigma^2 \left(\frac{1}{\sum w_i} + \frac{\bar{x}_w^2}{M} \right).$$

Det følger at

$$T_1 = \frac{\dfrac{\hat{\alpha} - \alpha}{\sqrt{\sigma^2 \left(\dfrac{1}{\sum w_i} + \dfrac{\bar{x}_w^2}{M} \right)}}}{\sqrt{\dfrac{\hat{S}}{\sigma^2(n-2)}}} = \frac{(\hat{\alpha} - \alpha)\sqrt{(n-2)}}{\sqrt{\hat{S} \left(\dfrac{1}{\sum w_i} + \dfrac{\bar{x}_w^2}{M} \right)}} \sim t_{\alpha/2}(n-2)$$

og

$$T_2 = \frac{\dfrac{\hat{\beta} - \beta}{\sqrt{\dfrac{\sigma^2}{M}}}}{\sqrt{\dfrac{\hat{S}}{\sigma^2(n-2)}}} = \frac{\sqrt{(n-2)M}(\hat{\beta} - \beta)}{\sqrt{\hat{S}}} \sim t_{\alpha/2}(n-2).$$

Jeg minner om at $M = \sum w_i(x_i - \bar{x}_w)^2$. Hvis alle responsvariablene har lik varians, ser vi at testobservatorene reduseres til uttrykkene for vanlig lineær regresjon.

9.2.2 Faremomenter ved bruk av vektet regresjon

Mange av faremomentene ved enkel, lineær regresjon kan generaliseres til enkel, lineær vektet regresjon. Jeg velger derfor ikke å gå i detalj. Likevel er det verdt å presisere, nok en gang, at vektet regresjon tillater heteroskedastisitet.

Som ved lineær regresjon kan problemer forekomme om $x_1, x_2, \ldots, x_n$ måles med feil. En tilleggsfaktor ved vektet regresjon er at også vektene, w_i, antas å være kjente størrelser, målt uten feil. Effekten av å nytte estimerte vekter er vanskelig å vurdere. Erfaringer viser at liten variasjon i vektene grunnet estimering ofte ikke påvirker resultatene i særlig grad. Likevel kan effekten være betydelig og uforutsigbar om vektene estimeres fra et lite antall replikerte observasjoner [39]. Simuleringsresultater senere i oppgaven vil vise effekten av å bruke vektet regresjon når de estimerte vektene er korrelerte med de estimerte verdiene av responsvariabelen.

10 Regresjonsbaserte tester for å identifisere publikasjonsbias

I dette kapittelet vil jeg gi en innføring i noen få, kjente tester for publikasjonsbias basert på regresjon. I tillegg vil jeg innføre en ny testmetode. Utfordringer knyttet til de ulike testmetodene vil kort diskuteres. Disse kan variere i ulike situasjoner, og jeg vil først klargjøre scenarioet jeg ønsker å undersøke testmetodene under.

10.1 Beskrivelse av metaanalysesituasjon

I Begg og Mazumdars simuleringssituasjon [8] er effektestimatene normalfordelte. Dette er rimelig fordi effektestimatet i flere tilfeller kan antas å ha en asymptotisk normalfordeling. Det nye scenarioet er ikke like generelt. Situasjonen er hentet fra artikkelen til Macaskill et al. [35]. Her har vi en konkret statistisk modell, hvor 2×2-tabeller genereres. Hver enkelt studie består av en behandlingsgruppe og en kontrollgruppe. Kun to utfall er mulige, suksess og fiasko. Effektestimatet måles i log-odds-ratio. Uten tap av generalitet antas den underliggende, sanne log-odds-ratioen mindre enn eller lik null. Dette tilsvarer odds-ratio mindre enn eller lik 1 og beskriver en behandling med ingen eller positiv effekt. Dette er motsatt av Begg og Mazumdars definisjoner [8], hvor en positiv effekt har en estimert verdi av odds-ratio som er større enn 1.

Vi har k studier. Hver studie har sampelstørrelse n_i, $i = 1, 2, \ldots, k$. Antall observerte suksesser og fiaskoer i behandlingsgruppen betegnes ved a_i og c_i henholdsvis. Tilsvarende er b_i og d_i antall observerte suksesser og fiaskoer i kontrollgruppen. Det følger at $n_i = a_i + b_i + c_i + d_i$. Behandlingseffekten estimeres ved $t_i = \ln\big((a_i d_i)/(b_i c_i)\big)$. Log-odds-ratio konvergerer raskere mot normalfordeling enn odds-ratio, grunnet den additive strukturen til logaritmen. Variansen til t_i estimeres ved $v_i = 1/a_i + 1/b_i + 1/c_i + 1/d_i$, selv om dette er et estimert asymptotisk resultat. Dette variansuttrykket utledes i Vedlegg B. Vi antar videre en modell med faste effekter. Det vil si at den underliggende behandlingseffekten er den samme for alle studier. En metaanalyse av våre k studier estimerer en verdi, $\hat{\delta}$, for den sanne effekten δ. Vi estimerer $\hat{\delta}$ ved hjelp av Mantel-Haenszels log-odds-ratio, som defineres [4] ved

$$\hat{\delta}_{MH} = \ln\left(\frac{\sum_{i=1}^{k} a_i d_i / n_i}{\sum_{i=1}^{k} b_i c_i / n_i}\right). \tag{4}$$

10.2 Eggers regresjonsmetoder

10.2.1 Metoder

I likhet med rangkorrelasjonsmetodene for å avdekke publikasjonsbias i metaanalyser er også Eggers regresjonsmetoder, introdusert av Egger et al. [15], basert på et empirisk grunnlag. Testene tar utgangspunkt i hvordan funnelplottet forandrer form dersom metaanalysen inneholder publikasjonsbias. Jeg vil først presentere metoden. Motivasjonen følger i neste seksjon.

Egger et al. [15] måler asymmetri i funnelplottet ved hjelp av en lineær regresjonsanalyse. Jeg kaller metoden EU, i likhet med Macaskill et al. [35]. I en metaanalyse med k studier observerer vi studienes effektestimater, $t_1, t_2, \ldots, t_k$, og de tilhørende variansestimatene, $v_1, v_2, \ldots, v_k$. Regresjonslinjen uttrykkes ved

$$z_i = \alpha + \frac{\beta}{\sqrt{v_i}},$$

hvor det standardiserte effektestimatet, $z_i = t_i/\sqrt{v_i}$, er den avhengige variabelen. Effektestimatets presisjon, $1/\sqrt{v_i}$, er den uavhengige variabelen. Metoden tilsvarer en enkel, lineær regresjon av punktene i et Galbraith-plott [19], men regresjonslinjen går ikke nødvendigvis gjennom origo.

Skjæringspunktet, α, defineres. Dernest testes nullhypotesen,

$$H_0 : \ \alpha = 0,$$

mot en passende ensidig eller tosidig alternativ hypotese. Forkastning av denne nullhypotesen medfører også forkastning av nullhypotesen

$$H_0 : \text{ Ingen publikasjonsbias.}$$

Egger et al. [15] introduserer også vektet lineær regresjon som en metode for å teste for publikasjonsbias. Den avhengige og uavhengige variabelen er definert som ved EU. Vektene er definert ved $w_i = 1/v_i$. Metoden vil jeg kalle EW.

10.2.2 Motivasjon

Vi har et Galbraith-plott, hvor studienes estimerte presisjon avsettes langs den horisontale aksen, og studienes standardiserte effektestimat avsettes langs den vertikale aksen. Funnelplottet baseres på det faktum at presisjonen ved estimering av den underliggende behandlingseffekten øker når sampelstørrelsen til studiene i metaanalysen øker [15]. Små studier har gjerne lav presisjon. De vil derfor ligge nær null langs den horisontale aksen i Galbraith-plottet. Små studier vil også ligge nær null langs den vertikale aksen, relativt til større studier med tilsvarende

verdier for effektestimatene. Dette er fordi effektestimatene standardiseres ved å multiplisere dem med deres presisjon.

Større studier vil ha høyere presisjon. De vil derfor bevege seg bort fra origo langs den horisontale aksen i figuren hvor de standardiserte effektestimatene plottes mot presisjonen. Det samme er tilfellet langs den vertikale aksen. Disse studiene har høye absolutte standardiserte effektestimater, sammenliknet med små studier med tilsvarende effektestimater. Punktene i et Galbraith plott uten publikasjonsbias vil derfor ligge spredt rundt en lineær linje som går gjennom origo. Stigningstallet indikerer både størrelsen og retningen på den sanne, underliggende effekten [15].

Eggers intuitive resonnement under nullhypotesen kan vises matematisk. Under nullhypotesen har vi at

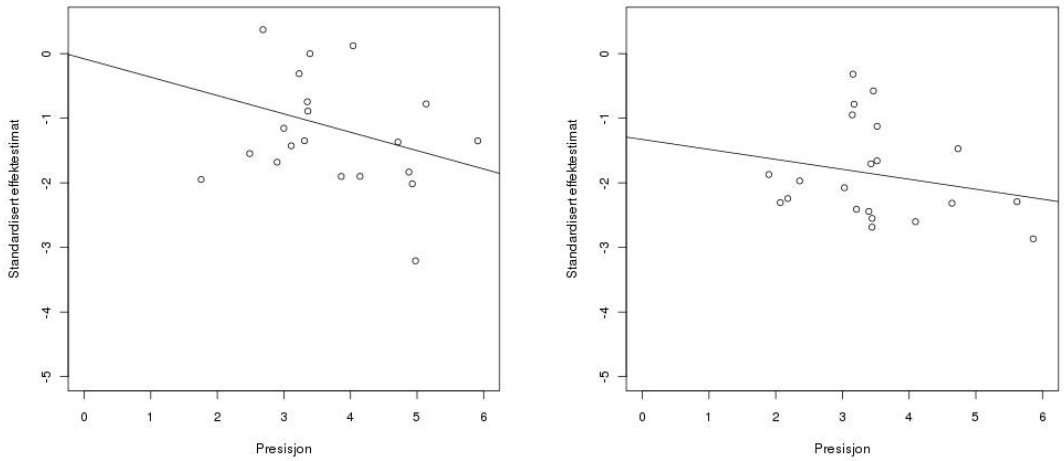
$$\mathrm{E}(z_i|v_i) = \mathrm{E}\left(\frac{t_i}{\sqrt{v_i}}|v_i\right) = \frac{1}{\sqrt{v_i}}\mathrm{E}(t_i|v_i) = \beta\frac{1}{\sqrt{v_i}},$$

hvor β er stigningstallet til regresjonslinjen. Linjen går gjennom origo, og t_i er en forventningsrett estimator for stigningstallet. Igjen tillater jeg misbruk av notasjon. Her betegner v_i den virkelige innsatte verdien av variansen til t_i. I andre situasjoner bruker jeg gjerne v_i som notasjon for den estimerte variansen. Jeg savner et matematisk resonnement hos Egger et al. [15].

Dersom det finnes mange små studier som viser motsatt effekt av de store studiene, vil disse muligens ikke bli publisert. Vi får da asymmetri i funnelplottet. Den aktuelle regresjonslinjen vil ikke lenger gå gjennom origo. Skjæringspunktet, α, kan derfor brukes som et mål for asymmetri. Jo mer α avviker fra null, desto mer markert vil asymmetrien være [15].

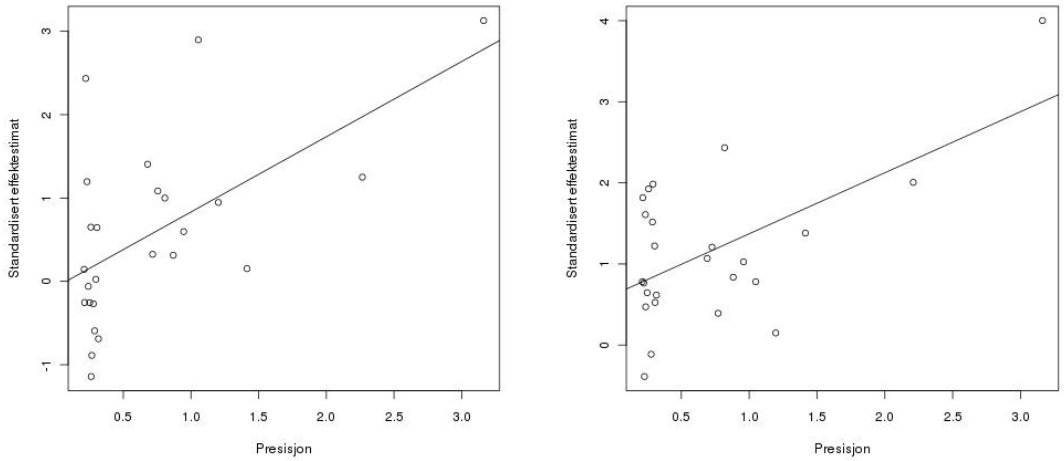
Egger et al. [15] bruker den naturlige logaritmen til odds-ratio som mål på effektstørrelse. Effekten er beskyttende dersom den estimerte verdien til odds-ratio er mindre enn 1. Regresjonslinjen vil ligge under origo dersom små studier viser beskyttende effekter. Negative verdier av skjæringspunktet indikerer at små studier viser mer utpregede gunstige effekter enn store studier [15]. Figur 5 viser dette. Med Begg og Mazumdars definisjon av odds-ratio, vil regresjonslinjen ligge over origo dersom små studier viser beskyttende effekt. Her indikerer positive verdier av skjæringspunktet at små studier viser tydeligere gunstige effekter enn store studier, se Figur 6.

Siden effektestimatene standardiseres, vil alle Y_i ha tilnærmet de samme variansene asymptotisk. Likevel er det en nærliggende tanke at ikke alle de standardiserte effektestimatene bør vektlegges like mye. I situasjoner hvor det er flere små studier, men bare én stor, påstår Egger et al. [15] en kan oppnå bedre styrke ved å vekte analysen med vektene $w_i = 1/v_i$.



(a) Uten publikasjonsbias. (b) Med publikasjonsbias.

Figur 5: De standardiserte effektestimatene plottes mot effektestimatenes presisjon. Eksempel med og uten publikasjonsbias basert på simulerte metaanalyser. Odds ratio mindre enn 1 indikerer beskyttende effekter.



(a) Uten publikasjonsbias. (b) Med publikasjonsbias.

Figur 6: De standardiserte effektestimatene plottes mot effektestimatenes presisjon. Eksempel med og uten publikasjonsbias basert på simulerte metaanalyser. Odds ratio større enn 1 indikerer beskyttende effekter.

10.2.3 Utfordringer knyttet til Eggers regresjonsmetoder

Som tidligere nevnt er det en konvensjonell oppfatning at de k stokastiske parene (t_1, v_1), $(t_2, v_2), \ldots, (t_k, v_k)$ fra studiene i en metaanalyse er uavhengige og har den samme kontinuerlige bivariate fordelingen. Responsvariablene, $t_i/\sqrt{v_i}$, $i = 1, 2, \ldots, k$, vil derfor være uavhengige.

Asymptotisk, når sampelstørrelsene i de ulike studiene er store, vil de betingede standardiserte effektestimatene, gitt de virkelige verdiene av variansene, være normalfordelte. Da vil også

$$\mathrm{Var}(z_i|v_i) = \mathrm{Var}\left(\frac{t_i}{\sqrt{v_i}}\Big|v_i\right) = \frac{1}{v_i}\mathrm{Var}(t_i|v_i) = \frac{v_i}{v_i} = 1,$$

hvor v_i er den innsatte virkelige verdien av variansen til t_i. I testsituasjonen til Macaskill et al. vil forutsetningen om homoskedastisitet kun være tilnærmet oppfylt asymptotisk. Det er fordi vi estimerer variansen ved $1/a_i + 1/b_i + 1/c_i + 1/d_i$. Ved utføring av inferens om α, kan t-fordeling nyttes selv om feilleddene bare er tilnærmet normalfordelte. Konklusjonene vil ikke påvirkes av moderate avvik fra normalfordelingen [26], særlig ikke dersom metaanalysen inneholder mange studier.

Den uavhengige variabelen, $1/\sqrt{v_i}$, inneholder målefeil fordi variansene er estimert ut fra de observerte dataene. For binære data vil den estimerte variansen være et biased estimat av den virkelige variansen. Skjevheten minker når studienes sampelstørrelser øker [4]. Målefeil i forklaringsvariabelen kan føre til et estimert stigningstall som er biased. Skjevhetens utstrekning og retning avhenger av variansen og kovariansen til den sanne verdien av den uavhengige variabelen og målefeilen, som vist i forrige kapittel. Det estimerte stigningstallet er biased mot null om det ikke er korrelasjon mellom målefeilen og den virkelige verdien av den uavhengige variabelen. Dette vil igjen medføre skjevhet i estimatet for skjæringspunktet. Hvis vi antar beskyttende effekter i form av odds-ratio mindre enn 1, kan vi forvente et negativt skjæringspunkt, selv om metaanalysen ikke inneholder publikasjonsbias [35]. En metaanalyse inneholder et endelig antall studier. Dersom variansene feilestimeres, kan dette muligens medføre noe ekstra heteroskedastisitet i regresjonsmodellen. Ved vektet regresjon kan problemene forringes ytterligere, da også vektene er målt med feil.

I testsituasjonen til Macaskill et al. er variansen en funksjon av den estimerte log-odds-ratio [35]. Den estimerte standardiserte behandlingseffekten er korrelert med dens estimerte presisjon [24]. Hvis, for en gitt studie, den observerte odds-ratioen ligger nærmere 1 enn den sanne, underliggende verdien, vil denne observasjonens vekt, $1/v_i$, ofte estimeres for høyt [35]. Tilsvarende vil variansene ofte estimeres for lavt. Altså risikerer vi at estimeringsfeilen i effektestimatet forplanter seg videre ved estimering av variansen. Vi er strengt tatt ikke i en situasjon hvor mitt tidligere bevis for fortynningsskjevhet gjelder. I beviset, gitt i Seksjon 9.1.3, forutsatte jeg at feilen i responsvariabelen var uavhengig av målefeilen i forklaringsvariabelen.

Det er temmelig uvanlig å vekte observasjoner ulikt i situasjoner hvor det allerede er sørget for homoskedastisitet. Eggers vektede regresjonsmetode passer således ikke overens med vektet regresjon, slik jeg har behandlet denne. Dette får konsekvenser for størrelsene på testobservatorene i utvalget. Macaskill et al. [35] beskriver konsekvenser av å nytte Eggers vektede regresjonsmetode på metaanalyser hvor det er én eller flere store studier. Her vil studiene med stor sampelstørrelse ha sterk innflytelse ved fastsetting av stigningstallet, noe som resulterer i økt variabilitet for skjæringspunktet.

En lineær modell vil være passende under nullhypotesen, $H_0 : \alpha = 0$, fordi $\mathrm{E}(z_i|v_i) = \beta/\sqrt{v_i}$. Modellen er også korrekt spesifisert. Siden forklaringsvariablene i reelle metaanalyser er stokastiske, kan man argumentere for at det uansett ikke vil forekomme problemer om den lineære modellen er dårlig spesifisert. Denne argumentasjonen er gitt i Seksjon 9.1.3.

Simuleringsresultater vil vise hvordan de ulike problemene vil vekselvirke og påvirke nivåestimatene.

10.3 Funnelplottregresjon

10.3.1 Tidligere introduserte metoder basert på funnelplottregresjon

Macaskill et al. [35] foreslår alternative testmetoder til Eggers regresjonsprosedyrer. Også disse nytter regresjonsanalyse. Metodene bruker de observerte dataene i et funnelplott direkte. Effektestimatet, t_i, er responsvariabelen. Studiestørrelsen, n_i, er forklaringsvariabelen. Her slipper man å estimere den uavhengige variabelen og problemene dette medfører.

Seksjon 1.3 omhandler funnelplott. Seksjonen gir den intuitive motivasjonen for funnelplottregresjon. Uten publikasjonsbias har regresjonslinjen et forventet stigningstall lik null. Dersom stigningstallet er signifikant ulik null, indikerer dette en korrelasjon mellom effektestimatene og studiestørrelsen, muligens grunnet publikasjonsbias. Nullhypotesen defineres ved

$$H_0 : \beta = 0.$$

Variansen til t_i er ikke nødvendigvis lik variansen til t_j, $i \neq j$. Det er nødvendig å nytte vektet regresjon for å tillate heteroskedastisitet. Den ene testmetoden bruker invers varians som sine vekter og kalles FIV. Denne vektingen er noe uheldig. Vektene er basert på de observerte dataene og er en funksjon av de estimerte log-odds-ratioene. En tendenserer å vekte studier med observert odds-ratio nærmere 1 enn den underliggende, sanne verdien, for høyt. I simuleringssituasjonene til Macaskill et al. risikerer regresjonslinjen å få et negativt stigningstall, selv uten publikasjonsbias [35].

Mackaskill et al. [35] forsøker derfor en alternativ vektet regresjonsmetode, FPV. Et "pooled" estimat av suksessproporsjonen i studie i er gitt ved $(a_i + b_i)/n_i$. Variansen er gitt

ved $1/(a_i + b_i) + 1/(c_i + d_i)$. Inversen til dette utrykket brukes som vekter i FPV. Forfatterne mener dette bør redusere korrelasjonen mellom vektene og den avhengige variabelen. Vektene er derimot estimert ut fra de observerte dataene. De vil fortsatt være utsatt for stokastisk variabilitet.

10.3.2 Ny metode basert på funnelplottregresjon

For ytterligere å redusere problemet med at vektene estimeres med feil som avhenger av de estimerte effektene, kan en nytte de observerte studienes sampelstørrelser som vekter. Sampelstørrelsene er ikke estimert ut fra de observerte dataene. En unngår dermed vekter som er utsatt for stokastisk variabilitet. Jeg innfører denne metoden og kaller den FN.

Vi husker at $\text{Var}(Y_i) = \sigma^2/w_i$. Hvis sampelstørrelsen er en biased tilnærming til w_i, kan dette følgelig gi problemer. Denne mulige utfordringen kan derimot heller ikke utelukkes ved de andre testmetodene basert på funnelplottregresjon. Eksempelvis vet man at den estimerte presisjonen er et biased estimat av den virkelige presisjonen, med større skjevhet for mindre sampelstørrelser [35]. Dette medfører at også vektene i FIV vil gi et biased estimat for w_i. Metoden min er vel verdt et forsøk.

11 Ny simuleringssituasjon, simuleringsresultater og diskusjon

Jeg ønsker å undersøke og sammenlikne testenes egenskaper. Macaskill et al. [35] sin metaanalysesituasjon er passende til formålet. En beskrivelse av denne metaanalysesituasjonen er gitt i Seksjon 10.1. En fordel med dette scenarioet er at jeg får sammenliknet Begg og Mazumdars testprosedyre [8] og den ustandardiserte testmetoden i situasjoner hvor variansfordelingen er mer reell, og hvor en ikke nødvendigvis har normalfordelte effektestimat. Resultater fra tidligere simuleringer viser et behov for dette. Jeg beskriver simuleringsprosedyren til Macaskill et al. [35]. Enkelte steder gjør jeg små endringer. Dette kommenteres underveis.

11.1 Simuleringer

11.1.1 Seleksjonsmodeller

Jeg har tidligere begrunnet at jeg ønsker å avgrense simuleringene til ensidig seleksjon basert på p-verdien for hypotesen om at den sanne, underliggende effekten er lik null. Dette vil jeg gjøre også her. Resultatene til Macaskill et al. [35] understøtter valget om å se bort fra tosidig seleksjon basert på p-verdi. Ingen av testmetodene presterer å avdekke den store proporsjonen av studier som ekskluderes fra metaanalysene når det ikke er noen underliggende effekt. Symmetrien i funnelplottene opprettholdes fordi de ekskluderte studiene for det meste befinner seg nær sentrum av funnelplottet [35]. Forfatterne utfører heller ikke simuleringer hvor seleksjonen bare avhenger av det observerte effektestimatet. Vektfunksjonen gis ved $w_i(p_i) = \exp(-\beta p_i^{\alpha})$, hvor $p_i = \Phi(t_i/\sqrt{v_i})$. Dette er den samme vektfunksjonen brukt av Begg og Mazumdar [8].

11.1.2 Parametervalg

Hver simulerte metaanalyse inneholder 21 studier. To ulike sammensetninger av studienes sampelstørrelser brukes. Konfigurasjon A inneholder 11 studier med 100 personer (testgjenstander) i både behandlingsgruppen og kontrollgruppen, seks studier med 200 personer per gruppe og fire studier med 300 personer per gruppe. Konfigurasjon B har ti studier med 100 personer i hver gruppe, fem med 200 per gruppe, tre med 300 per gruppe, to med 500 i hver gruppe og en studie med 1000 personer i både behandlingsgruppen og kontrollgruppen.

De ulike behandlingseffektene blir satt til $\ln(1)$, $\ln(2/3)$, $\ln(1/2)$ og $\ln(1/4)$, altså odds-ratio lik 1, 2/3, 1/2 og 1/4.

Vektfunksjonen med parameterverdiene $\alpha = 1.5$ og $\beta = 4$ gir sterk seleksjonbias. Hvert scenario repeteres med og uten publikasjonsbias.

Macaskill et al. ønsker også å sammenlikne testmetodene dersom antall studier per metaanalyse dobles eller tredobles. I tillegg vil forfatterne undersøke testenes styrke dersom studie-

nes sampelstørrelsene dobles eller tredobles, mens antall studier i metaanalysen fortsatt er 21. Her begrenser de seg til ensidig ekskludering av studier og til underliggende behandlingseffekter på ln(1) og ln(2/3). På bakgrunn av deres resultater velger jeg å begrense simuleringene ytterligere. Jeg tredobler antall studier per metaanalyse. I tillegg dobler jeg studienes sampelstørrelser, samtidig som antall studier per metaanalyse forblir 21. Jeg velger, i motsetning til forfatterne, å teste for behandlingseffekter lik ln(1), ln(2/3), ln(1/2) og ln(1/4). Konfigurasjon C og D inneholder 63 studier per metaanalyse, men tilsvarer ellers konfigurasjon A og B henholdsvis. Konfigurasjon E og F tilsvarer også konfigurasjon A og B, med unntak av at studienes sampelstørrelse er doblet.

11.1.3 Utføring

Jeg ønsker å generere metaanalyser med en gitt sammensetning av de inkluderte studienes sampelstørrelser. Jeg starter med å generere effektstørrelse og varians for en første studie med en av de gitte sampelstørrelsene. Den underliggende suksessannsynligheten i kontrollgruppen finnes ved tilfeldig trekning fra en uniform kontinuerlig fordeling på intervallet [0.1, 0.5]. Ved hjelp av verdien til den underliggende odds-ratioen, samt suksessannsynligheten i kontrollgruppen, kan vi beregne suksessannsynligheten i behandlingsgruppen. De simulerte verdiene av a_i og b_i finnes ved tilfeldig trekning fra en binomisk fordeling med suksessannsynligheten i henholdsvis behandlingsgruppen og kontrollgruppen. Vi definerer c_i som differansen mellom antall personer i behandlingsgruppen og a_i. Videre er d_i differansen mellom antall personer i kontrollgruppen og b_i.

Log-odds-ratio og varians estimeres ved hjelp av

$$t_i = \ln\left(\frac{(a_i + 0.5)(d_i + 0.5)}{(b_i + 0.5)(c_i + 0.5)}\right)$$

og

$$v_i = \frac{1}{a_i + 0.5} + \frac{1}{b_i + 0.5} + \frac{1}{c_i + 0.5} + \frac{1}{d_i + 0.5}.$$

Dette reduserer biasen i den estimerte log-odds-ratioen og forbedrer estimatoren for variansen [4].

Om studien publiseres eller ikke, avgjøres ved å trekke tilfeldig fra en bernoullifordeling med parameter w, hvor w er verdien til vektfunksjonen. Prosedyren ovenfor gjentas for hver av de gitte sampelstørrelsene til det ønskede antallet studier med disse sampelstørrelsene er valgt. Slik sikrer vi at den simulerte metaanalysen inneholder like mange studier som det metaanalyser typisk inneholder i virkeligheten, samt at vi får en rimelig fordeling av studiestørrelser. Mantel-Haenszels log-odds-ratio, $\hat{\delta}_{MH}$, estimeres på bakgrunn av studiene i den genererte metaanalysen. Denne estimatoren korrigeres ikke ved å addere 0.5 til hver celle, da denne er mer robust. Estimatoren er definert i Likning (4), Seksjon 10.1.

For hver simulerte metaanalyse utfører Macaskill et al. [35] alle testmetodene. Her avviker jeg noe. Jeg velger å utføre én testmetode for hver simulerte metaanalyse, slik at estimatene er uavhengige. Jeg utfører Eggers regresjonsmetode, både uvektet, EU, og vektet, EW. Jeg ønsker også å teste metodene basert på funnelplottregresjon, FIV, FPV og FN. Videre undersøker jeg egenskapene til Begg og Mazumdars rangkorrelasjonstest [8], BVS, beskrevet i Kapittel 3. De standardiserte effektestimatene korreleres mot effektestimatenes varians. Jeg tester også den ustandardiserte testen, BVU, hvor effektestimatene korreleres mot deres varianser. En alternativ testmetode undersøker rangkorrelasjonen mellom de standardiserte effektestimatene og studienes sampelstørrelse. Denne metoden betegnes BNS. Jeg utfører også den ustandardiserte versjonen til denne testen, kalt BNU. Effektestimatene korreleres mot studienes sampelstørrelse.

Prosessen repeteres 10000 ganger per testmetode. Nye varianser beregnes for hver replikasjon. Simuleringsituasjonen til Macaskill et al. [35] gjør det dermed vanskelig å implementere prosedyren som korrigerer nivået til Begg og Mazumdars testmetode. Den korrigerte testen er basert på den simulerte fordelingen til Kendalls tau, betinget på de estimerte variansene. Denne testen er beskrevet i Kapittel 8.

Det empiriske signifikansnivået beregnes når metaanalysen ikke er påvirket av publikasjonsbias. Den empiriske styrken beregnes når publikasjonsbias påvirker den simulerte metaanalysen. Både ensidige og tosidige tester utføres ved nominelle nivå på henholdsvis 0.05 og 0.10. Ved testing for publikasjonsbias i metaanalyser vil ensidige tester være passende dersom en forventer positiv effekt og dersom publikasjonsbias kommer av at studier som viser ingen eller liten effekt ikke inkluderes i metaanalysen [35].

Biasen og andelen av de genererte studiene som inkluderes i metaanalysene beregnes også. Biasen defineres ved $\mathrm{E}(\hat{\delta}_{MH} - \delta)$. Fordi Macaskill et al. [35] utfører alle testmetodene for hver simulerte metaanalyse, får forfatterne kun ett estimat for bias og andel studier som inkluderes i metaanalysen. Jeg får ett estimat per testmetode. Jeg ønsker ikke å inkludere alle disse estimatene i tabellene. Tabellene vil inneholde mye informasjon og bli uoversiktlige. En mulighet er å referere den gjennomsnittlige biasen og den gjennomsnittlige andelen inkluderte studier fra alle testmetodene. Ulempen er at disse estimatene må oppdateres om jeg ønsker å undersøke nye testmetoder. Estimatene for bias og andelen inkluderte studier avhenger av de ulike simuleringsscenarioene, ikke av testprosedyrene. Jeg gjentar derfor simuleringssituasjonene ovenfor, men uten å utføre tester. Jeg refererer biasen og andelen inkluderte studier fra disse uavhengige simuleringene. Dette vil gi representative estimat.

Forventning og standardfeil (SE) estimeres for skjæringspunktet for Eggers regresjonsmetoder i tilfeller hvor metaanalysene ikke inneholder publikasjonsbias. Likeledes estimeres forventning og standardfeil for stigningstallet for funnelplottmetodene. Vi kan da undersøke

om estimatorene, som testene er basert på, er biased ved å teste om de aktuelle forventningene avviker fra null.

Jeg utfører alle simuleringene i statistikkprogrammet R [43], i motsetning til Macaskill et al. [35] som bruker SAS. R-dokumentasjonen til funksjonen `cor.test()` gir ingen god beskrivelse av hvordan ties behandles. Det gjør derimot R-dokumentasjonen til funksjonen `Kendall()` [36]. Vi unngår ikke ties i simuleringssituasjonene til Macaskill et al. [35]. Videre vil Kendalls tau beregnes ved hjelp av `Kendall()`. I situasjoner hvor ties ikke forekommer, beregner funksjonen p-verdier for Kendalls tau under nullhypotesen om ingen publikasjonsbias ved å nytte en eksakt algoritme gitt av Best og Gipps [9]. Regresjonsanalysene utføres ved hjelp av funksjonen `summary.lm()` [43].

11.2 Simuleringsresultater

11.2.1 Simuleringsresultater for metaanalyser uten publikasjonsbias, konfigurasjon A og B

Hvis estimatorene som de ulike testene er basert på er forventningsrette, skal skjæringspunktets forventning være lik null for Eggers regresjonsmetoder i tilfeller hvor metaanalysene ikke er påvirket av publikasjonsbias. For funnelplottmetodene skal stigningstallets forventning være lik null. Tabell 28 presenterer disse tallene for konfigurasjon A og B, samt den maksimale standardfeilen over de ulike behandlingseffektene. Tabellen viser om forventningene avviker signifikant fra null ved et tosidig nivå på 0.05.

Ingen av testmetodenes estimatorer er biased under konfigurasjon A, når den underliggende, sanne log-odds-ratioen, δ, er lik null. EW, FPV og FN viser signifikant skjevhet under konfigurasjon B.

Når $\delta \neq 0$, avviker forventningen signifikant fra null for Eggers uvektede regresjonsmetode, EU. Det gjennomsnittlige skjæringspunktet er negativt. Skjevheten øker desto lenger bort fra nullverdien δ beveger seg. Dette er også tilfellet for EW, men fortegnet på skjevheten avhenger av konfigurasjonen. Stigningstallet til FPV og FN er ikke biased når den sanne effekten er ulik null. FIV har et gjennomsnittlig negativt stigningstall. Skjevheten øker jo lenger δ beveger seg bort fra null.

Nivåestimatene finnes i Tabell 29. Jeg ønsker å kontrollere nivåestimatene mot resultatene til Macaskill et al. [35]. Det er rimelig å påstå at resultatene i stor grad samsvarer, selv om en grundig sammenlikning er vanskelig. Forfatterne presenterer kun nivåestimatene visuelt. Jeg tar utgangspunkt i mine resultater.

Ensidige tester Først vurderes resultatene fra de ensidige testene. La p_1 være nivået til en testmetode. Tabell 30 viser nivåestimater når den underliggende log-odds-ratioen er lik null. I tillegg inkluderer tabellen tosidige konfidensintervaller med konfidensnivå 0.95 for p_1. For å kunne konstruere mer nøyaktige konfidensintervaller, presenteres nivåestimatene med flere desimaler enn i Tabell 29.

EU og FIV har signifikansnivå som kan antas å være lik det nominelle når $\delta = 0$. Det gjelder også FPV og FN, selv om resultatene i Tabell 28 indikerer bias for konfigurasjon B. Nivået til BVS, BNS og BNU ligger for lavt. Vi forkaster nullhypotesen om at det virkelige nivået er lik det nominelle for BVU under konfigurasjon A. Nivået ligger for høyt. Denne nullhypotesen forkastes ikke under konfigurasjon B. EW presterer dårlig med tanke på nivå. Dette er spesielt tydelig ved konfigurasjon B.

Når den underliggende, sanne effekten beveger seg bort fra nullverdien, er det kun FPV og FN som generelt har et nivå som tilsvarer det nominelle. Tosidige konfidensintervaller med konfidensnivå 0.95, viser derimot at vi forkaster nullhypotesen om at disse metodene har et ensidig signifikansnivå som er lik det nominelle når $\delta = \ln(1/4)$. Disse konfidensintervallene inkluderes ikke her. EU, EW, BVS og BVU har stort sett ensidige nivå som ligger godt over 0.05. Simuleringsresultatene i Tabell 29 viser videre at FIV, BN og BNU har et nivå som ligger under det nominelle.

Tosidige tester Jeg beveger meg bort fra den ensidige testen og vurderer den tosidige testen med et nominelt nivå på 0.10. Det empiriske tosidige signifikansnivået finnes ved å summere estimatene under *ensidig* og *andre hale* i Tabell 29. Under nullhypotesen forventer jeg at 5% av de genererte metaanalysene skal vise statistisk signifikant publikasjonsbias i hver retning. Dersom vi sammenlikner estimatene under *ensidig* og *andre hale*, ser vi at dette, i de fleste situasjoner, ikke stemmer. Når den underliggende effekten er null, vil vi stort sett forkaste like mange metaanalyser i hver retning, men ikke nødvendigvis 5%. Denne symmetrien avtar for flere testmetoder jo lenger bort fra nullverdien δ beveger seg. EU, BVU og BVS viser den tydeligste asymmetrien. FPV og FN viser ikke markant asymmetri. Tosidige konfidensintervaller med konfidensnivå 0.95 gir likevel forkastning av nullhypotesen om at like mange metaanalyser viser signifikant publikasjonsbias i hver retning under konfigurasjon A og B når $\delta = \ln(1/4)$, og under konfigurasjon B når $\delta = \ln(1/2)$.

Simuleringsresultatene viser at FPV og FN har tosidige signifikansnivå som tilsvarer det nominelle nivået. FIV har grovt sett tosidige signifikansnivå som i flere tilfeller ikke ligger langt unna 0.10. Det samme kan sies om EU og BVS. Disse metodene får derimot et nivå som er for høyt når den underliggende effekten ligger langt fra nullverdien samtidig som det er liten variasjon mellom studienes sampelstørrelse. Resultatene viser at EW har et tosidig nivå

som er omtrent doblet sammenliknet med det nominelle under konfigurasjon B. BVU har et tosidig nivå som hovedsaklig ligger noe høyere enn det nominelle. De resterende testmetodene, BNS og BNU, har generelt et nivå som ligger litt under 0.10.

11.2.2 Simuleringsresultater for metaanalyser med publikasjonsbias, konfigurasjon A og B

Tabell 31 viser simuleringsresultatene for metaanalyser med publikasjonsbias. Maksimal standardfeil for styrkeestimatene i denne tabellen er 0.005. Dette tallet er funnet ved funksjonsdrøfting av uttrykket for standardfeilen, jamfør Vedlegg A og og funksjonsdrøftingen i Seksjon 4.1.1. Mine resultater samsvarer jevnt over med resultatene til Macaskill et al. [35]. De avviker derimot noe for BNS (som Macaskill et al. betegner BN). Det er vanskelig å gi en forklaring på avvikene. Jeg har ikke informasjon om hvordan Macaskill et al. beregner p-verdien til tau under nullhypotesen om ingen publikasjonsbias. Hvordan behandler forfatterne ties?

Jeg vil kort presentere de viktigste trekkene ved mine simuleringsresultater. Færre studier ekskluderes fra metaanalysene når den underliggende effekten beveger seg bort fra nullverdien. Problemet med publikasjonsbias minker derfor med økende absolutte verdier av δ [35], noe den estimerte biasen understøtter. Testmetodenes tosidige styrke synker deretter.

Testenes styrke avhenger også av konfigurasjonen. Jo mer studienes sampelstørrelser varierer, desto lettere vil metodene stort sett identifisere publikasjonsbias.

Den generelle styrken er lav, selv med et tosidig nominelt nivå på 0.10. Dette er også tilfellet selv om nesten to tredeler av studiene ekskluderes fra metaanalysene. EU, BVU og BVS viser hovedsaklig de samme tendensene og gir best styrke. EW gir noe svakere styrkeestimater. Funnelplottregresjonsmetodene ser ut til å prestere nokså likt med tanke på styrke. Disse testene gir styrkeestimater som ligger en god del lavere enn estimatene til EU, BVS, BVU og EW. De svakeste styrkeestimatene finnes hos BNS og BNU.

11.2.3 Effekten av å øke antall studier per metaanalyse, konfigurasjon C og D

Tabell 32 viser de gjennomsnittlige skjæringspunktene for Eggers testmetoder og de gjennomsnittlige stigningstallene for funnelplottmetodene. Ingen av estimatorene, som de ulike testene er basert på, har et forventet skjæringspunkt eller stigningstall som avviker signifikant fra null når $\delta = 0$. Resultatene viser at estimatorene er biased når $\delta \neq 0$. Unntaket er FPV og FN.

Nivået til de ulike testene påvirkes minimalt av å øke antall studier per metaanalyse når det ikke er noen underliggende effekt. Disse resultatene presenteres i Tabell 33. FPV og FN viser fortsatt ingen markant asymmetri når δ beveger seg bort fra nullverdien. I de fleste tilfeller har de et signifikansnivå som er nær det nominelle, både ved ensidige og tosidige tester.

Asymmetrien er derimot mer markant for de resterende testmetodene under konfigurasjon C og D, enn den var under henholdsvis konfigurasjon A og B. EW har, under konfigurasjon D, fortsatt et signifikansnivå som er omtrent dobbelt så høyt som det nominelle.

Bedre styrke og mer informasjon oppnås ofte ved å øke sampelstørrelsen. Testmetodene for publikasjonsbias er intet unntak. Resultatene i Tabell 34 for $\delta = 0$ bekrefter dette. Samtlige tester oppnår hovedsaklig bedre styrke under konfigurasjon C og D, sammenliknet med henholdsvis konfigurasjon A og B. Tendensene er ellers de samme. Styrken synker desto lenger δ beveger seg bort fra nullverdien. Stor variasjon i studiers sampelstørrelse bedrer også styrken. Resultatene mine ser ut til å sammenfalle med resultatene presentert av Macaskill et al [35]. Fortsatt viser EU best styrke, etterfølgt av BVU og BVS. Styrkeestimatene avviker ikke stort blant de tre funnelplottmetodene. Konfidensintervaller med tosidig konfidensnivå 0.95 viser at FN i enkelte tilfeller vil være å foretrekke framfor FPV. FIV har generelt høyest sannsynlighet for å gjøre feil av type II.

11.2.4 Effekten av å øke studienes sampelstørrelse, konfigurasjon E og F

Vi beveger oss over til konfigurasjon E og F. Her dobles sampelstørrelsen til de ulike studiene i metaanalysen. Tabell 35 viser at kun skjæringspunktet til EU under konfigurasjon F inneholder signifikant bias når det ikke er noen underliggende effekt. Når $\delta \neq 0$, vil estimatorene til EU, EW og FIV være biased. Skjevheten øker jo lenger bort fra nullverdien den underliggende effekten beveger seg. Stigningstallet til FPV og FN viser ingen signifikant bias.

Tabell 36 viser nivåeestimatene. Disse gir ikke mye ny informasjon, sammenliknet med nivåestimatene hvor studienes sampelstørrelse er halvert. Rangkorrelasjonsmetodene, med unntak av BVU, har fortsatt både ensidige og tosidige signifikansnivå som er lavere enn de nominelle når $\delta = 0$. De resterende testmetodene, ikke medregnet EW, har grovt sett ensidige og tosidige nivå som kan antas lik de nominelle i disse tilfellene. Det er verdt å merke seg at asymmetrien i de ulike testene under konfigurasjon E og F generelt ser ut til å minke når $\delta \neq 0$, sammenliknet med henholdsvis konfigurasjon A og B.

Macaskill et al. [35] beskriver en moderat økning i styrke for å avdekke publikasjonsbias ved dobling av studiestørrelsen når det ikke er er noen underliggende effekt. Ved en tredobling refererer forfatterne kun en liten styrkeøkning sammenliknet med hva de fikk ved en dobling. Macaskill et al. [35] forklarer denne lille økningen med at styrken er begrenset av antall studier i metaanalysen, grunnet liten variasjon innad i hver enkelt studie. Denne antakelsen er rimelig på et generelt grunnlag.

Mine resultater viser, grovt sett, ingen endring i styrke ved en dobling av sampelstørrelsen til metaanalysens studier når $\delta = 0$, sammenliknet med konfigurasjon A og B. Disse resultatene finnes i Tabell 37. Dette står i sterk kontrast til resultatene presentert av Macaskill et al. [35],

beskrevet i avsnittet over.

Vil en endring i studienes sampelstørrelse kunne medføre en endring i andelen genererte studier som inkluderes i metaanalysene når $\delta = 0$? Vektfunksjonen avhenger av p-verdien for hypotesen om ingen underliggende effekt. Under nullhypotesen skal p-verdien være uniformt fordelt på intervallet [0,1]. I simuleringsscenarioet til Macaskill et al. [35] defineres p-verdien som $\Phi(t_i/\sqrt{v_i})$. Det vil si at testobservatoren, $t_i/\sqrt{v_i}$, antas standardnormalfordelt. Asymptotisk er dette en rimelig antakelse. For et endelig antall forsøkspersoner per studie kan vi forvente skjevhet i estimatet for p-verdien. Denne skjevheten vil minke om studienes sampelstørrelse økes, da testobservatorens tilnærming til standardnormalfordelingen bedres. Simuleringsresultatene viser derimot at andelen studier som inkluderes i metaanalysene holder seg omtrent konstant ved en dobling av studienes sampelstørrelse når det ikke er noen underliggende effekt. Seleksjonen vil således ikke forårsake endring i styrkeestimatene.

Fordi seleksjonen ikke påvirkes, medfører en økning av sampelstørrelsene per studie ellers kun en reskalering av funnelplottet når $\delta = 0$. Testmetodene er skaleringsinvariante. Det er rimelig at styrken ikke endres i dette tilfellet. Dette støtter opp om mine simuleringsresultater.

Studienes sampelstørrelse er større under konfigurasjon E og F enn under konfigurasjon A og B henholdsvis. Flere av de genererte studiene vil derfor inkluderes i metaanalysene når den underliggende effekten beveger seg bort fra nullverdien. Som et resultat av dette, synker testmetodenes styrke. Her er mine simuleringsresultater konsistente med de få resultatene Macaskill et al. [35] rapporterer.

Tabell 28: Gjennomsnittlig skjæringspunkt (Eggers regresjonsmetoder) og gjennomsnittlig stigningstall (funnelplottregresjon) for de simulerte metaanalysene uten publikasjonsbias. Konfigurasjon A og B.

Underliggende log-odds-ratio (δ)	Konfigurasjon	Eggermetoden (skjæringspunkt)		Funnelplottregresjon (stigningstall)		
		EU	EW	FIV	FPV	FN
ln(1)						
	A	-5.9×10^{-3}	-7.7×10^{-3}	4.4×10^{-6}	4.4×10^{-6}	6.2×10^{-6}
	B	4.7×10^{-3}	$-1.7 \times 10^{-2*}$	-8.3×10^{-7}	$1.7 \times 10^{-6*}$	$1.6 \times 10^{-6*}$
ln(2/3)						
	A	$-1.0 \times 10^{-1*}$	$-4.6 \times 10^{-2*}$	$-2.1 \times 10^{-5*}$	-4.2×10^{-6}	-2.0×10^{-7}
	B	$-2.8 \times 10^{-2*}$	2.4×10^{-2}	$-5.5 \times 10^{-6*}$	-3.3×10^{-7}	-2.0×10^{-7}
ln(1/2)						
	A	$-2.0 \times 10^{-1*}$	$-1.0 \times 10^{-1*}$	$-2.8 \times 10^{-5*}$	4.4×10^{-6}	3.0×10^{-6}
	B	$-5.0 \times 10^{-2*}$	$4.6 \times 10^{-2*}$	$-8.8 \times 10^{-6*}$	-3.4×10^{-7}	-2.6×10^{-7}
ln(1/4)						
	A	$-3.5 \times 10^{-1*}$	$-1.4 \times 10^{-1*}$	$-1.0 \times 10^{-4*}$	-3.6×10^{-6}	4.5×10^{-6}
	B	$-9.9 \times 10^{-2*}$	$9.9 \times 10^{-1*}$	$-2.3 \times 10^{-5*}$	6.2×10^{-6}	9.3×10^{-7}
Maksimal SE						
	A	8.7×10^{-3}	9.6×10^{-3}	4.1×10^{-4}	4.3×10^{-6}	4.8×10^{-6}
	B	5.8×10^{-3}	7.5×10^{-3}	9.4×10^{-7}	9.6×10^{-7}	1.0×10^{-6}

Maksimal SE = maksimal standardfeil for parameterestimatene.

* Estimatorene inneholder statistisk signifikant bias ($|z| = |\text{gjennomsnitt}|/\text{SE} > z_{\alpha/2} = 1.96$).

Konfigurasjon A (21 studier: 11 × 100/gruppe, 6 × 200/gruppe, 4 × 300/gruppe).

Konfigurasjon B (21 studier: 10 × 100/gruppe, 5 × 200/gruppe, 3 × 300/ gruppe, 2 × 500/ gruppe, 1 × 1000/gruppe).

Tabell 29: Signifikansnivå for de forskjellige testmetodene ved konfigurasjon A og B. Resultater for ensidige tester på 5% nivå og tosidige tester på 10% nivå (sum av ensidig og andre hale) vises.

Behandlings-effekt (δ)	Konfigu-rasjon	Bias	% inkluderte studier	Eggermetoden Nivå %				Funnelplottregresjon Nivå %								Rangkorrelasjon Nivå %					
				EU		EW		FIV		FPV		FN		BVS		BVU		BNS		BNU	
				en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale
ln(1)																					
	A	-0.000	100	**5.3**	5.1	**5.9**	5.5	**5.3**	5.0	**5.4**	5.0	**5.1**	4.9	**4.5**	5.0	**5.6**	5.3	**4.1**	4.0	**4.3**	4.2
	B	0.000	100	**5.1**	4.9	**10.9**	10.7	**5.1**	5.5	**5.2**	5.0	**5.3**	5.0	**4.4**	4.1	**5.1**	4.5	**4.0**	3.8	**4.0**	4.1
ln(2/3)																					
	A	-0.001	100	**6.9**	3.4	**6.2**	4.7	**4.4**	5.6	**4.8**	5.2	**4.8**	5.0	**6.1**	3.2	**7.8**	3.9	**3.7**	4.5	**3.8**	4.2
	B	0.000	100	**5.8**	4.7	**10.5**	11.1	**4.5**	5.6	**5.2**	5.0	**5.2**	5.1	**4.9**	3.8	**6.5**	4.5	**3.6**	4.0	**3.9**	4.7
ln(1/2)																					
	A	-0.000	100	**8.7**	2.9	**7.6**	4.5	**4.3**	5.7	**4.9**	5.0	**4.8**	5.1	**7.9**	2.4	**9.3**	2.9	**3.6**	4.3	**3.5**	4.3
	B	-0.000	100	**6.1**	4.0	**9.4**	11.3	**4.1**	5.9	**4.6**	5.4	**4.6**	5.6	**6.0**	2.8	**8.1**	3.2	**3.3**	4.8	**3.6**	4.1
ln(1/4)																					
	A	0.000	100	**12.6**	1.6	**8.7**	3.4	**3.3**	7.5	**4.4**	5.2	**4.5**	5.5	**11.7**	1.3	**17.0**	1.3	**2.7**	5.6	**3.1**	5.0
	B	-0.000	100	**7.4**	2.8	**9.0**	13.2	**3.6**	8.3	**4.4**	5.4	**4.5**	5.5	**8.5**	1.7	**12.4**	2.1	**2.4**	6.0	**3.1**	4.1

Konfigurasjon A (21 studier: 11 × 100/gruppe, 6 × 200/gruppe, 4 × 300/gruppe).
Konfigurasjon B (21 studier: 10 × 100/gruppe, 5 × 200/gruppe, 3 × 300/ gruppe, 2 × 500/ gruppe, 1 × 1000/gruppe).

Tabell 30: Nivåestimat og tosidige konfidensintervaller for p_1 med konfidensnivå 0.95, $\delta = 0$.

Testmetode	Konfigurasjon	Estimat for p_1	Konfidensintervall for p_1
EU			
	A	0.0526	[0.04822, 0.05698]
	B	0.0505	[0.04621, 0.05479]
EW			
	A	0.0585	[0.05390, 0.06310]
	B	0.1090	[0.10290, 0.11511]
FIV			
	A	0.0533	[0.04890, 0.05770]
	B	0.0507	[0.04640, 0.05500]
FPV			
	A	0.0539	[0.04947, 0.05833]
	B	0.0519	[0.04755, 0.05625]
FN			
	A	0.0510	[0.04669, 0.05531]
	B	0.0530	[0.04861, 0.05739]
BVS			
	A	0.0447	[0.04065, 0.04875]
	B	0.0438	[0.03979, 0.04781]
BVU			
	A	0.0556	[0.05111, 0.06009]
	B	0.0513	[0.04698, 0.05562]
BNS			
	A	0.0411	[0.03721, 0.04499]
	B	0.0395	[0.03568, 0.04332]
BNU			
	A	0.0429	[0.03893, 0.04687]
	B	0.0402	[0.03635, 0.04405]

Tabell 31: Styrke for å oppdage publikasjonsbias for de forskjellige testmetodene ved konfigurasjon A og B. Resultater for ensidige tester på 5% nivå og tosidige tester på 10% nivå (sum av ensidig og andre hale) vises.

Behandlings-effekt (δ)	Konfigu-rasjon	Bias	% inkluderte studier	Eggermetoden Styrke %				Funnelplottregresjon Styrke %						Rangkorrelasjon Styrke %					
				EU		EW		FIV		FPV		FN		BVS		BVU		BNS		BNU	
				en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale
ln(1)																					
	A	-0.199	36.6	**35.8**	0.1	**31.6**	0.3	**25.7**	0.4	**26.4**	0.4	**26.7**	0.5	**32.3**	0.2	**35.2**	0.2	**23.3**	0.3	**22.4**	0.4
	B	-0.160	36.5	**58.2**	0.0	**55.7**	0.4	**41.0**	0.2	**40.7**	0.1	**41.4**	0.2	**42.7**	0.0	**46.2**	0.0	**36.2**	0.1	**36.6**	0.1
ln(2/3)																					
	A	-0.065	79.2	**28.6**	0.4	**23.1**	1.1	**16.9**	1.2	**18.0**	1.2	**18.3**	0.9	**24.8**	0.4	**26.9**	0.5	**14.8**	0.7	**14.1**	0.9
	B	-0.043	80.8	**31.1**	0.4	**28.9**	3.4	**18.7**	1.5	**19.6**	1.4	**19.9**	1.4	**26.4**	0.2	**28.8**	0.3	**16.7**	0.4	**18.2**	0.4
ln(1/2)																					
	A	-0.025	92.8	**19.1**	1.0	**14.2**	2.1	**8.5**	3.0	**9.7**	2.5	**10.3**	2.4	**15.8**	0.7	**18.5**	1.1	**6.7**	2.1	**7.8**	1.8
	B	-0.016	93.4	**13.7**	1.5	**15.8**	8.1	**7.9**	4.2	**8.7**	3.4	**9.9**	3.4	**14.0**	0.7	**16.9**	1.0	**7.3**	1.6	**7.7**	1.7
ln(1/4)																					
	A	-0.005	99.3	**15.5**	1.2	**9.8**	3.3	**3.3**	6.7	**4.5**	4.7	**4.9**	4.8	**12.8**	1.0	**17.6**	1.3	**3.1**	4.7	**3.9**	4.8
	B	-0.002	99.3	**9.1**	2.5	**8.6**	13.2	**3.3**	8.1	**4.5**	5.5	**4.8**	5.8	**9.4**	1.1	**14.0**	1.5	**2.6**	5.2	**3.4**	4.6

Konfigurasjon A (21 studier: 11 × 100/gruppe, 6 × 200/gruppe, 4 × 300/gruppe).
Konfigurasjon B (21 studier: 10 × 100/gruppe, 5 × 200/gruppe, 3 × 300/ gruppe, 2 × 500/ gruppe, 1 × 1000/gruppe).

Tabell 32: Gjennomsnittlig skjæringspunkt (Eggers regresjonsmetoder) og gjennomsnittlig stigningstall (funnelplottregresjon) for de simulerte metaanalysene uten publikasjonsbias. Konfigurasjon C og D.

Underliggende log-odds-ratio (δ)	Konfigurasjon	Eggermetoden (skjæringspunkt)		Funnelplottregresjon (stigningstall)		
		EU	EW	FIV	FPV	FN
ln(1)						
	C	-4.8×10^{-3}	-9.6×10^{-3}	-3.5×10^{-7}	-2.6×10^{-6}	6.4×10^{-7}
	D	2.4×10^{-3}	1.4×10^{-3}	-7.9×10^{-7}	6.6×10^{-7}	-6.3×10^{-7}
ln(2/3)						
	C	$-1.0 \times 10^{-1*}$	$-5.3 \times 10^{-2*}$	$-1.6 \times 10^{-5*}$	-1.8×10^{-6}	-7.1×10^{-8}
	D	-2.8×10^{-2}	$2.4 \times 10^{-2*}$	$-4.6 \times 10^{-6*}$	-2.4×10^{-7}	-1.8×10^{-7}
ln(1/2)						
	C	$-1.9 \times 10^{-1*}$	$-1.0 \times 10^{-1*}$	$-3.4 \times 10^{-5*}$	3.1×10^{-6}	-2.6×10^{-6}
	D	$-4.7 \times 10^{-2*}$	$3.7 \times 10^{-2*}$	$-7.8 \times 10^{-6*}$	3.0×10^{-7}	2.4×10^{-7}
ln(1/4)						
	C	$-3.9 \times 10^{-1*}$	$-1.8 \times 10^{-1*}$	$-9.9 \times 10^{-5*}$	1.1×10^{-6}	1.1×10^{-6}
	D	$-1.2 \times 10^{-1*}$	$9.0 \times 10^{-2*}$	$-2.4 \times 10^{-5*}$	1.4×10^{-8}	2.3×10^{-7}
Maksimal SE						
	C	4.9×10^{-3}	5.5×10^{-3}	3.1×10^{-4}	2.4×10^{-6}	2.7×10^{-6}
	D	3.3×10^{-3}	4.5×10^{-3}	5.1×10^{-7}	5.3×10^{-7}	5.8×10^{-7}

Maksimal SE = maksimal standardfeil for parameterestimatene.

* Estimatorene inneholder statistisk signifikant bias ($|z| = |\text{gjennomsnitt}|/\text{SE} > z_{\alpha/2} = 1.96$).

Konfigurasjon C (63 studier: 33 × 100/gruppe, 18 × 200/gruppe, 12 × 300/gruppe).

Konfigurasjon D (21 studier: 30 × 100/gruppe, 15 × 200/gruppe, 9 × 300/ gruppe, 6 × 500/ gruppe, 3 × 1000/gruppe).

Tabell 33: Signifikansnivå for de forskjellige testmetodene ved konfigurasjon C og D. Resultater for ensidige tester på 5% nivå og tosidige tester på 10% nivå (sum av ensidig og andre hale) vises.

Behandlings-effekt (δ)	Konfigu-rasjon	Bias	% inkluderte studier	Eggermetoden Nivå %				Funnelplottregresjon Nivå %								Rangkorrelasjon Nivå %					
				EU		EW		FIV		FPV		FN		BVS		BVU		BNS		BNU	
				en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale
ln(1)																					
	C	-0.000	100	**5.3**	5.0	**5.7**	5.2	**4.6**	4.5	**4.9**	5.2	**5.0**	5.2	**5.0**	4.6	**5.5**	5.6	**4.4**	4.6	**4.4**	4.6
	D	0.000	100	**5.2**	5.2	**11.1**	11.3	**5.0**	5.2	**5.1**	4.7	**5.0**	5.2	**4.5**	4.8	**5.5**	5.3	**4.8**	4.2	**4.1**	4.1
ln(2/3)																					
	C	-0.000	100	**8.2**	3.2	**6.9**	4.6	**4.0**	5.6	**5.0**	5.3	**5.1**	5.1	**8.4**	2.5	**10.1**	2.3	**3.8**	4.8	**3.9**	5.6
	D	0.000	100	**5.8**	4.1	**11.1**	12.5	**4.2**	6.6	**4.5**	5.2	**4.8**	5.4	**6.7**	2.5	**8.7**	3.6	**3.4**	4.6	**3.9**	4.7
ln(1/2)																					
	C	-0.000	100	**10.5**	1.8	**8.8**	3.9	**3.4**	7.0	**4.7**	4.7	**4.8**	5.2	**12.0**	1.5	**14.6**	1.7	**3.4**	5.6	**3.7**	5.5
	D	-0.000	100	**7.2**	3.9	**10.5**	13.3	**3.7**	7.1	**5.0**	5.0	**5.1**	5.1	**8.6**	1.8	**11.9**	2.4	**2.6**	5.4	**3.9**	5.4
ln(1/4)																					
	C	-0.001	100	**25.3**	1.2	**12.0**	2.8	**1.9**	10.4	**4.2**	4.8	**4.6**	5.3	**23.9**	0.3	**33.3**	0.3	**2.2**	8.4	**2.1**	8.5
	D	0.000	100	**11.1**	1.6	**8.6**	8.8	**1.8**	11.9	**4.4**	5.2	**4.8**	5.3	**14.5**	0.7	**23.6**	0.8	**1.6**	8.5	**2.8**	6.3

Konfigurasjon C (63 studier: 33 × 100/gruppe, 18 × 200/gruppe, 12 × 300/gruppe).
Konfigurasjon D (21 studier: 30 × 100/gruppe, 15 × 200/gruppe, 9 × 300/ gruppe, 6 × 500/ gruppe, 3 × 1000/gruppe).

Tabell 34: Styrke for å oppdage publikasjonsbias for de forskjellige testmetodene ved konfigurasjon C og D. Resultater for ensidige tester på 5% nivå og tosidige tester på 10% nivå (sum av ensidig og andre hale) vises.

Behandlings-effekt (δ)	Konfigu-rasjon	Bias	% inkluderte studier	Eggermetoden Styrke %				Funnelplottregresjon Styrke %						Rangkorrelasjon Styrke %					
				EU		EW		FIV		FPV		FN		BVS		BVU		BNS		BNU	
				en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale
ln(1)																					
	C	-0.198	35.9	**72.7**	0.0	**63.4**	0.0	**54.3**	0.0	**55.6**	0.0	**55.7**	0.1	**72.0**	0.0	**72.9**	0.0	**54.2**	0.0	**54.5**	0.0
	D	-0.165	35.9	**93.8**	0.0	**87.3**	0.0	**80.6**	0.0	**81.8**	0.0	81.4	0.0	**85.8**	0.0	**87.9**	0.0	**78.5**	0.0	**79.1**	0.0
ln(2/3)																					
	C	-0.066	78.6	**60.7**	0.0	**46.4**	0.1	**33.2**	0.3	**36.5**	0.2	**37.6**	0.2	**58.1**	0.1	**58.7**	0.0	**32.5**	0.1	**32.5**	0.2
	D	-0.045	80.2	**64.9**	0.0	**45.5**	1.2	**35.3**	0.6	**38.2**	0.4	**40.9**	0.4	**62.0**	0.0	**62.8**	0.0	**41.6**	0.1	**40.9**	0.1
ln(1/2)																					
	C	-0.024	92.5	**39.1**	0.1	**24.3**	0.8	**12.3**	3.0	**15.3**	1.2	**16.7**	1.3	**36.1**	0.1	**39.9**	0.2	**12.2**	1.3	**12.1**	1.3
	D	-0.016	93.1	**27.8**	0.5	**18.9**	7.2	**10.0**	3.2	**12.6**	2.1	**14.4**	2.0	**29.8**	0.1	**34.4**	0.2	**11.5**	1.0	**12.8**	1.0
ln(1/4)																					
	C	-0.003	99.2	**29.0**	0.3	**13.3**	2.2	**2.5**	9.1	**5.2**	4.2	**6.1**	4.3	**27.5**	0.2	**37.8**	0.3	**3.4**	6.7	**2.7**	6.7
	D	-0.003	99.3	**12.9**	1.6	**9.7**	16.3	**2.1**	11.3	**4.8**	4.7	**5.1**	5.0	**17.9**	0.4	**26.5**	0.5	**2.0**	7.5	**3.5**	0.5

Konfigurasjon C (63 studier: 33 × 100/gruppe, 18 × 200/gruppe, 12 × 300/gruppe).
Konfigurasjon D (21 studier: 30 × 100/gruppe, 15 × 200/gruppe, 9 × 300/ gruppe, 6 × 500/ gruppe, 3 × 1000/gruppe).

Tabell 35: Gjennomsnittlig skjæringspunkt (Eggers regresjonsmetoder) og gjennomsnittlig stigningstall (funnelplottregresjon) for de simulerte metaanalysene uten publikasjonsbias. Konfigurasjon E og F.

Underliggende log-odds-ratio (δ)	Konfigurasjon	Eggermetoden (skjæringspunkt)		Funnelplottregresjon (stigningstall)		
		EU	EW	FIV	FPV	FN
ln(1)						
	E	-9.8×10^{-5}	1.3×10^{-3}	-2.7×10^{-7}	-2.1×10^{-6}	-6.0×10^{-8}
	F	$-1.2 \times 10^{-2*}$	3.6×10^{-3}	2.6×10^{-8}	2.2×10^{-7}	-1.7×10^{-7}
ln(2/3)						
	E	$-5.6 \times 10^{-2*}$	-1.7×10^{-2}	$-4.4 \times 10^{-6*}$	-7.9×10^{-7}	-8.7×10^{-7}
	F	$-3.0 \times 10^{-2*}$	$2.0 \times 10^{-2*}$	$-9.4 \times 10^{-7*}$	2.3×10^{-7}	-3.2×10^{-7}
ln(1/2)						
	E	$-1.3 \times 10^{-1*}$	$-6.1 \times 10^{-2*}$	$-1.0 \times 10^{-5*}$	5.2×10^{-7}	-2.4×10^{-7}
	F	$-3.0 \times 10^{-2*}$	$2.2 \times 10^{-2*}$	$-2.3 \times 10^{-6*}$	1.9×10^{-7}	-2.1×10^{-7}
ln(1/4)						
	E	$-2.7 \times 10^{-1*}$	$-1.1 \times 10^{-1*}$	$-2.6 \times 10^{-5*}$	2.6×10^{-8}	1.6×10^{-6}
	F	$-7.8 \times 10^{-2*}$	$7.8 \times 10^{-2*}$	$-6.2 \times 10^{-6*}$	-1.0×10^{-8}	5.4×10^{-8}
Maksimal SE						
	E	8.6×10^{-3}	9.6×10^{-3}	1.5×10^{-6}	1.5×10^{-6}	1.7×10^{-6}
	F	5.8×10^{-3}	7.4×10^{-3}	3.4×10^{-7}	3.4×10^{-7}	3.5×10^{-7}

Maksimal SE = maksimal standardfeil for parameterestimatene.

* Estimatorene inneholder statistisk signifikant bias ($|z| = |\text{gjennomsnitt}|/\text{SE} > z_{\alpha/2} = 1.96$).

Konfigurasjon E (21 studier: 11 × 200/gruppe, 6 × 400/gruppe, 4 × 600/gruppe).

Konfigurasjon F (21 studier: 10 × 400/gruppe, 5 × 400/gruppe, 3 × 600/ gruppe, 2 × 1000/ gruppe, 1 × 2000/gruppe).

Tabell 36: Signifikansnivå for de forskjellige testmetodene ved konfigurasjon E og F. Resultater for ensidige tester på 5%-nivå og tosidige tester på 10%-nivå (sum av ensidig og andre hale) vises.

Behandlings-effekt (δ)	Konfigu-rasjon	Bias	% inkluderte studier	Eggermetoden Nivå %				Funnelplottregresjon Nivå %								Rangkorrelasjon Nivå %					
				EU		EW		FIV		FPV		FN		BVS		BVU		BNS		BNU	
				en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale
ln(1)																					
	E	0.000	100	**5.0**	5.1	**5.4**	5.6	**4.8**	4.8	**4.7**	5.0	**5.2**	5.2	**4.5**	4.4	**5.1**	5.4	**3.9**	4.0	**4.0**	4.3
	F	0.000	100	**5.1**	4.9	**10.0**	10.5	**4.6**	5.2	**5.0**	4.9	**5.3**	5.0	**4.3**	4.7	**5.0**	5.5	**3.7**	3.8	**3.9**	3.8
ln(2/3)																					
	E	0.000	100	**5.9**	4.8	**6.2**	5.8	**4.7**	5.9	**5.3**	4.9	**5.0**	5.0	**5.6**	3.3	**6.4**	3.6	**3.9**	4.1	**3.9**	4.5
	F	0.000	100	**5.8**	4.6	**10.0**	10.8	**4.7**	5.4	**4.9**	5.0	**5.0**	5.1	**4.9**	3.5	**6.1**	4.5	**3.5**	3.7	**3.7**	3.9
ln(1/2)																					
	E	-0.001	100	**7.0**	3.4	**6.8**	4.8	**3.9**	5.4	**4.9**	5.0	**5.0**	5.1	**6.9**	3.4	**7.7**	3.5	**3.9**	4.3	**4.1**	4.4
	F	-0.000	100	**5.7**	4.5	**10.4**	10.9	**4.4**	6.0	**5.1**	5.5	**4.9**	5.3	**5.9**	3.3	**6.7**	3.8	**3.3**	4.6	**3.9**	3.9
ln(1/4)																					
	E	0.000	100	**10.9**	2.2	**8.3**	4.3	**3.6**	6.5	**4.6**	5.1	**4.5**	5.3	**9.7**	1.8	**12.3**	1.9	**3.0**	5.4	**3.3**	5.5
	F	-0.000	100	**7.0**	3.5	**8.9**	12.0	**3.7**	7.0	**4.7**	5.4	**4.5**	5.4	**7.3**	2.2	**10.2**	2.8	**2.9**	4.8	**3.3**	4.5

Konfigurasjon E (21 studier: 11 × 200/gruppe, 6 × 400/gruppe, 4 × 600/gruppe).

Konfigurasjon F (21 studier: 10 × 400/gruppe, 5 × 400/gruppe, 3 × 600/ gruppe, 2 × 1000/ gruppe, 1 × 2000/gruppe).

Tabell 37: Styrke for å oppdage publikasjonsbias for de forskjellige testmetodene ved konfigurasjon E og F. Resultater for ensidige tester på 5%-nivå og tosidige tester på 10%-nivå (sum av ensidig og andre hale) vises.

Behandlings-effekt (δ)	Konfigu-rasjon	Bias	% inkluderte studier	Eggermetoden Styrke %				Funnelplottregresjon Styrke %						Rangkorrelasjon Styrke %					
				EU		EW		FIV		FPV		FN		BVS		BVU		BNS		BNU	
				en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale	en-sidig	andre hale
ln(1)																					
	E	-0.140	36.6	**34.2**	0.2	**30.6**	0.5	**26.2**	0.5	**26.7**	0.4	**26.4**	0.4	**30.9**	0.2	**32.6**	0.2	**24.1**	0.3	**23.7**	0.3
	F	-0.113	36.7	**55.9**	0.0	**57.2**	0.3	**40.1**	0.2	**41.9**	0.1	**42.0**	0.2	**41.9**	0.1	**45.4**	0.0	**37.4**	0.1	**35.8**	0.1
ln(2/3)																					
	E	-0.024	89.4	**18.4**	0.8	**15.9**	1.5	**12.0**	2.0	**12.4**	1.8	**13.1**	1.8	**15.7**	0.7	**17.9**	1.1	**9.6**	1.2	**9.9**	1.4
	F	-0.015	90.4	**17.3**	0.9	**19.6**	6.1	**10.9**	2.8	**11.7**	2.4	**11.9**	2.8	**14.5**	0.6	**16.8**	0.9	**10.6**	0.8	**10.0**	1.1
ln(1/2)																					
	E	-0.005	98.0	**9.8**	2.2	**8.7**	4.1	**5.5**	4.6	**6.4**	3.6	**6.5**	3.7	**8.5**	1.7	**10.4**	2.1	**4.6**	3.1	**4.7**	3.4
	F	-0.003	98.1	**7.9**	2.8	**11.9**	10.2	**5.1**	5.0	**6.3**	4.7	**5.7**	4.7	**7.6**	1.7	**9.7**	2.2	**4.3**	3.4	**4.8**	3.1
ln(1/4)																					
	E	-0.0003	99.9	**10.5**	1.9	**8.6**	4.0	**3.4**	6.3	**4.9**	5.0	**4.7**	5.3	**10.1**	1.7	**13.3**	1.9	**3.2**	5.2	**3.1**	4.8
	F	-0.0006	99.9	**7.2**	3.4	**9.3**	11.5	**3.4**	7.1	**4.9**	4.8	**4.3**	5.8	**7.0**	2.1	**10.4**	2.5	**2.8**	5.3	**3.1**	4.5

Konfigurasjon E (21 studier: 11 × 200/gruppe, 6 × 400/gruppe, 4 × 600/gruppe).

Konfigurasjon F (21 studier: 10 × 400/gruppe, 5 × 400/gruppe, 3 × 600/ gruppe, 2 × 1000/ gruppe, 1 × 2000/gruppe).

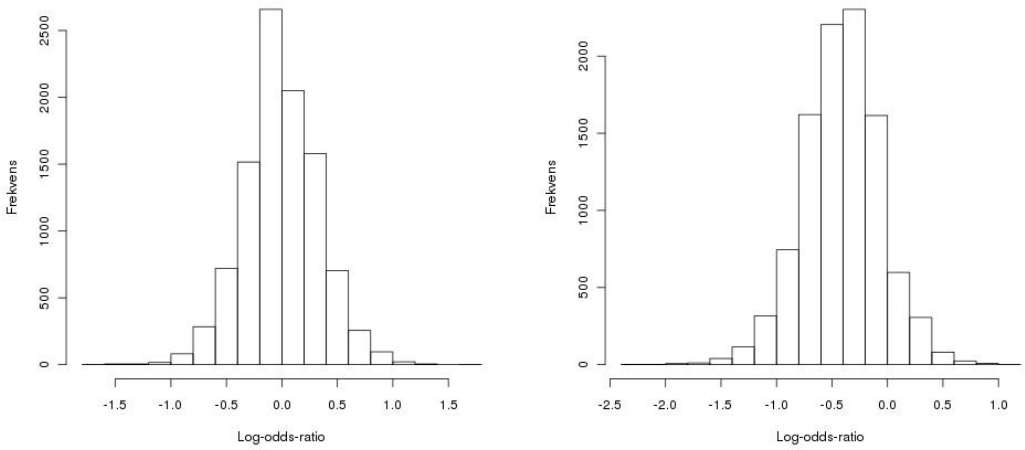
11.3 Asymmetri

En nullhypotese testes mot en tosidig alternativ hypotese ved et nominelt signifikansnivå α. Ved simulering under nullhypotesen forventer en å forkaste $(\alpha/2)100\%$ av de genererte metaanalysene i hver retning grunnet tilfeldigheter. Simuleringsresultatene viser derimot en sterk asymmetri for de fleste testmetodene. Asymmetrien ser ut til å avhenge av verdien til den underliggende, sanne log-odds-ratioen.

Begg [6] forklarer deler av denne asymmetrien ved fordelingen til log-odds-ratio. For et endelig antall forsøkspersoner per studie i metaanalysen, er denne fordelingen kun symmetrisk når forventningsverdien er null. Fordelingen til effektestimatet blir mer asymmetrisk desto lenger den sanne verdien av log-odds-ratio beveger seg bort fra nullverdien. Figur 7 viser denne skjevheten for ulike verdier av δ. Denne figuren er simulert med utgangspunkt i simuleringssituasjonen til Macaskill et al. [35], under nullhypotesen om ingen publikasjonsbias. Prosedyren til Macaskill et al. kjøres kun en gang, slik at vi får én enkelt metaanalyse. Denne metaanalysen inneholder 10000 studier. Hver av disse studiene har en kontrollgruppe og behandlingsgruppe bestående av 100 forsøkspersoner. Legg merke til forskjellene i venstre hale etterhvert som δ beveger seg bort fra nullverdien. Uavhengig av figurene, men ut fra den samme simuleringssituasjonen, beregner jeg skjevhetskoeffisienten til effektestimatet, $(1/n)\sum_{i=1}^{n}(t_i-\bar{t})^3/\left((1/n)\sum_{i=1}^{n}(t_i-\bar{t})^2\right)^{3/2}$. Når $\delta = \ln(1)$, kalkuleres den til 0.0166168. Når $\delta = \ln(1/4)$, er skjevhetskoeffisienten -0.665766. Beregningene bekrefter Beggs påstand.

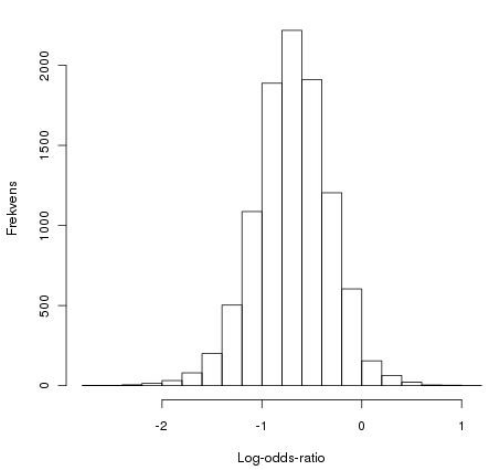
Beggs påstand underbygges videre av at asymmetrien under nullhypotesen ser ut til å være mindre ved konfigurasjon B enn ved konfigurasjon A. Flere studier har høyere sampelstørrelse under konfigurasjon B. Asymmetrien minker når jeg dobler sampelstørrelsene til studiene i metaanalysen, se Tabell 36. Denne mulige tendensen vil jeg undersøke nærmere. Jeg utfører nye simuleringer for de ulike testmetodene og tar utgangspunkt i konfigurasjon A. Jeg holder antall studier i metaanalysen konstant, men multipliserer studienes sampelstørrelse med faktoren 100. Effektestimatenes tilnærming til normalfordelingen vil bedres. Simuleringsresultatene, ikke vist i oppgaven, bekrefter at asymmetrien har minket betraktelig under nullhypotesen om ingen publikasjonsbias.

Alle testmetodene presentert i denne oppgaven er introdusert på et intuitivt grunnlag. En forventer at funnelplottet likner en omvendt symmetrisk trakt i tilfeller uten publikasjonsbias. I simuleringsscenarioet til Macaskill et al. [35] ser vi at funnelplottet kan være asymmetrisk selv i metaanalyser som ikke er utsatt for selektiv publikasjon. Simuleringsresultatene antyder at resultater fra de ulike testmetodene bør tolkes med forsiktighet dersom effektmålene i metaanalysen har en asymmetrisk fordeling i tilfeller uten publikasjonsbias. Funnelplottasymmetri er i slike situasjoner ikke et optimalt kriterium for å identifisere selektiv publikasjon.

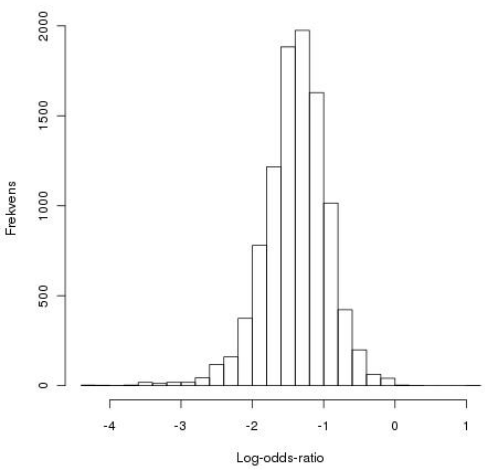


(a) ln(1)

(b) ln(2/3)



(c) ln(1/2)



(d) ln(1/4)

Figur 7: Histogram for effektestimatene.

11.4 Testmetodene basert på rangkorrelasjon

Oppgaven omhandler testmetoder for å identifisere publikasjonsbias i metaanalyser. Hoveddelen tar for seg testmetoder som baseres på rangkorrelasjon. Jeg ønsker kort å drøfte om resultatene fra Kapittel 4 og Kapittel 7 lar seg generalisere til situasjoner hvor variansene har en mer realistisk variansfordeling, og hvor en ikke nødvendigvis kan anta normalfordelte effektestimat.

Jeg har tidligere vist at rangkorrelasjonsmetodene ikke oppfyller forutsetningene for å utføre hypotesetester basert på Kendalls tau. I simuleringsscenarioet til Macaskill et al. [35] får vi ytterlige utfordringer. Problemer knyttet til den asymmetriske fordelingen til log-odds-ratio er behandlet i den forrige seksjonen. Av sammenlikningsårsaker har jeg gjennomført ensidige tester for BVS og BVU i simuleringssituasjonen til Begg og Mazumdar, beskrevet i Kapittel 3. Simuleringsresultatene bekrefter at asymmetrien ikke oppstår når effektestimatene er normalfordelte. Disse resultatene presenteres ikke i oppgaven.

Variansen estimeres. Ifølge Macaskill et al. er variansestimatet en funksjon av den estimerte verdien av log-odds-ratio [35]. Hvis, for en gitt studie, den observerte odds-ratioen ligger nærmere 1 enn den underliggende, sanne verdien, vil variansen ha en tendens til å bli estimert for lavt. Dette kan føre til en negativ korrelasjon mellom variansene og effektestimatene i situasjoner uten publikasjonsbias. Muligens kan dette forklare hvorfor det tosidige simulerte nivået til BVS generelt ser ut til å ligge noe høyere enn det estimerte signifikansnivået til BNS, i situasjoner hvor det ikke er noen underliggende effekt.

I denne seksjonen vurderer jeg konfigurasjonene A-F samlet. Nivået til BNS og BNU ligger hovedsaklig lavere enn det nominelle, både ved ensidig og tosidig testing, uansett verdien av den underliggende effekten. Det tosidige nivået til BNS og BNU ligger dessuten lavere enn det tosidige nivået til BVS og BVU for samtlige verdier av δ. Derimot viser BVS og BVU bedre tosidig styrke. I simuleringsscenarioet til Macaskill et al. klarer jeg ikke å kontrollere feil av type I på et gitt nivå. Når en tester for publikasjonsbias i metaanalyser, er det viktigere å begrense feil av type II enn feil av type I. BNS og BNU presterer dårligst ut fra en beslutningsteoretisk analyse. Jeg ønsker ikke å diskutere BNS og BNU mer inngående.

Simuleringsresultater fra Begg og Mazumdars testsituasjon [8] i Kapittel 8 viser grovt sett at den standardiserte og ustandardiserte testobservatoren inneholder omtrent den samme mengden informasjon. Om dette er tilfellet i det nye simuleringsscenarioet, er uvisst. Likevel vet vi at testmetodenes styrke må vurderes i sammenheng med nivået.

Når $\delta = 0$ og effektestimatene symmetrisk fordelt, viser de presenterte simuleringsresultatene at det tosidige nivået vil være lik det nominelle eller noe lavere for testmetoden introdusert av Begg og Mazumdar. Det tosidige nivået vil være lik det nominelle eller noe høyere for

den ustandardiserte testen. Den ustandardiserte testen oppnår bedre tosidig styrke. Når den underliggende effekten er ulik null, vil den ustandardiserte testen hovedsaklig ha lavest feil av type II. Samtidig viser dessverre denne testmetoden en mer markant asymmetri under nullhypotesen om ingen publikasjonsbias.

I tradisjonell statistikk er man opptatt av å begrense feil av type I fordi man tror og håper at dataene skal støtte opp om den alternative hypotesen. Dette argumentet kan ikke nyttes i vår nåværende situasjon. Vi forventer at dataene skal støtte opp om nullhypotesen. En bør ta de noe uvanlige rollene til nullhypotesen og den alternative hypotesen i betraktning, på samme måte som i Kapittel 7. Fra en beslutningsteoretisk analyse er det ikke urimelig å velge den ustandardiserte testmetoden framfor den standardiserte, dersom vi ikke har mulighet til å tilpasse nivået. Konklusjonen basert på resultatene fra metaanalysescenarioene til Macaskill et al. samsvarer med konklusjonen basert på Begg og Mazumdars testsituasjoner.

Under nullhypotesen om ingen publikasjonsbias, skal forventningen til Kendalls tau være lik null. Etterpåklokskap viser at jeg (og Macaskill et al.) burde inkludert tabeller som estimerer denne forventningen for de ulike rangkorrelasjonsmodellene, på lik linje som jeg har estimert eksempelvis forventningen til skjæringspunktet for EU. Slike resultater kan gi utfyllende informasjon om rangkorrelasjonstestene. Jeg overlater dette til videre arbeid.

11.5 Anbefaling av testmetode

Simuleringsresultatene i testsituasjonen til Macaskill et al. indikerer at testmetodenes styrke avhenger av ulike faktorer. Den underliggende behandlingseffekten, fordelingen av studienes sampelstørrelse og antall studier i metaanalysen vil påvirke styrkeestimatene. Testmetodene viser generelt lav styrke. Således er resultatene for BVS og BVU konsistente med styrkeestimatene fra Begg og Mazumdars simuleringssituasjon, hvor effektestimatene er normalfordelte. Lav styrke er ikke et stort problem når den underliggende behandlingseffekten ligger langt fra nullverdien. Selektiv publikasjon forekommer nærmest ikke i disse tilfellene. En større bekymring er de dårlige styrkeverdiene når metaanalysene inneholder få studier. Få studier er ikke uvanlig i praksis, særlig innen medisinsk forskning [8]. Denne innsikten er nyttig. Det er viktig ikke å utelukke publikasjonsbias i metaanalyser, selv om vi ikke forkaster nullhypotesen om ingen publikasjonsbias. Det er også viktig å legge merke til at styrken er lav, selv med et nominelt signifikansnivå på 0.10. Forsiktighet bør vises, uansett valg av testmetode.

Vanligvis vurderes og sammenliknes testmetodenens egenskaper ut fra sannsynligheten for å gjøre feil av type II. Dette forutsetter derimot at feil av type I kan kontrolleres på et gitt nivå. Dette lar seg ikke gjøre for flere av de ulike testmetodene. Testmetodenes styrke bør vurderes i sammenheng med nivået [35]. Høyere feil av type I vil generelt medføre lavere feil av type II, så sant en ikke endrer sampelstørrelsen eller konstruerer nye testmetoder [22]. EU,

BVU og BVS viser de beste styrkeestimatene. Dessverre antyder nivåestimatene generelt et nivå som ligger over det nominelle. BNS og BNU ser ut til å gi lavest styrke. Til gjengjeld har disse metodene også det laveste nivået. FPV har lavere styrke enn flere andre testmetoder. Fordi FPV generelt viser et tilnærmet korrekt nivå, foretrekker Macaskill et al. likevel denne testmetoden.

En test som forkaster H_0 med sannsynlighet 1, vil aldri kunne gjøre feil av type II. Tilsvarende vil en test som forkaster H_0 med sannsynlighet 0, alltid gjøre feil av type II. Det er viktig å kontrollere en feiltype, selv om en ikke har mulighet til å kontrollere den mest alvorlige. Macaskill et al. konkluderer i tråd med dette. Dette er ingen motsetning til argumentasjonen i Seksjon 11.4, hvor BVU anbefales framfor BVS. I denne situasjonen klarer ingen av testene å *kontrollere* noen av feiltypene på et gitt nivå. Det er stadig viktigere å *begrense* feil av type II enn feil av type I i en metaanalysesammenheng.

Min nye testmetode, FN, presterer hovedsaklig likt som FPV med tanke på nivå og styrke. FN oppnår i enkelte tilfeller noe lavere feil av type II. Vi unngår vekter som er utsatt for stokastisk variabilitet. FN vil være den foretrukne testen blant de ulike metodene basert på funnelplottregresjon. Jeg har således konstruert en forbedring til testmetodene introdusert av Macaskill et al. [35].

Vil testobservatoren til FPV og FN inneholde mer informasjon enn eksempelvis testobservatoren til EU, BVS eller BVU? Simuleringsresultatene viser at dette ikke er tilfellet når den underliggende log-odds-ratioen er null. Grovt sett har samtlige av disse testmetodene et signifikansnivå som tilsvarer det nominelle. FPV og FN gir de laveste styrkeestimatene. Disse testobservatorene inneholder således minst informasjon. EU oppnår generelt best styrke. Denne testmetoden foretrekkes hvis en ikke har noen underliggende effekt. Her har effektestimatene en symmetrisk fordeling under nullhypotesen om ingen publikasjonsbias. Om dette også vil være tilfellet når δ beveger seg bort fra nullverdien, er vanskelig å svare på uten å korrigere testmetodene slik at de kontrollerer feil av type I på et gitt nivå. Likevel kan det diskuteres om en anbefaling av FPV og FN framfor de resterende testmetodene er rimelig.

Den praktiske betydningen av å velge de ulike testmetodene bør tas i betraktning, særlig fordi en ikke lett kan avgjøre hvilke testobservatorer som inneholder mest informasjon når $\delta \neq 0$. En må ikke glemme de uvanlige rollene til nullhypotesen og den alternative hypotesen. Jeg henviser til argumentasjonen i Kapittel 7 om en føler behov for en videre oppfriskning rundt denne tematikken. Vi ønsker ikke å utføre metaanalyser hvor vi risikerer skjevhet i estimatet for den underliggende, sanne effekten. Macaskill et al. tar ikke hensyn til denne problematikken i sin vurdering av de ulike testmetodene.

Med utgangspunkt i en noe svevende og hypotetisk beslutningsteoretisk analyse kan EU, etterfulgt av BVU og BVS, anbefales. Disse testene oppnår langt bedre styrke enn FPV og

FN i tilfeller hvor biasen ikke er neglisjerbar. Dessverre forbedres styrken på bekostning av nivået. En anbefaling av Eggers uvektede regresjonsmetode for å identifisere publikasjonsbias i metaanalyser passer overens med eksisterende praksis. Et enkelt søk i databasen Web of Knowledge [45] viser at artikkelen til Egger et al. [15] er sitert i overkant av 5000 ganger. Begg og Mazumdars artikkel [8] er sitert nesten 1800 ganger, mens artikkelen til Macaskill et al. [35] er sitert i underkant av 300 ganger. Søket er utført 14. august 2012.

EW produserer uforutsigbare resultater, spesielt med tanke på nivået. Når metaanalysene inneholder 63 studier, konfigurasjon C og D, er styrken langt lavere for EW enn for EU, BVU og BVS. Samtidig er nivået i flere tilfeller høyere. EW har høyest sannsynlighet for å gjøre feil av både type I og type II. Denne metoden vil ikke anbefales.

Jeg vil ikke våge å gi en endelig anbefaling av testmetode. Valg av testmetode bør avhenge av hvordan hver enkelt stiller seg til hypotesetesting fundamentalt. Skal kontroll av feil av type I på et gitt nivå veie tyngst, eller skal en velge den observatoren som inneholder mest informasjon? Fordi testobservatoren i Eggers metode ser ut til å inneholde mer informasjon enn de andre observatorene, bør man arbeide videre med denne modellen. Det vil være ønskelig å tilpasse nivået til denne observatoren.

11.6 En enkel sammenlikning av Eggers regresjonsmetoder og de korrigerte rangkorrelasjonstestene i Begg og Mazumdars simuleringssituasjonen

Som tidligere nevnt er det ikke lett å implementere prosedyren som korrigerer nivået til rangkorrelasjonsmetodene i simuleringsscenarioet til Macaskill et al. [35]. Siden vi ikke klarer å kontrollere feil av type I på et gitt nivå, er det vanskelig å avgjøre hvilke testobservatorer som inneholder mest informasjon.

Jeg implementerer Eggers uvektede regresjonsmetode i testsituasjonen til Begg og Mazumdar [8], beskrevet i Kapittel 3. Som vanlig begrenses simuleringene til ensidig seleksjon basert på p-verdien for hypotesen om at den underliggende effekten er lik null. Simuleringsprosedyren gjentas 10000 ganger. Det tosidige nominelle nivået settes til 0.05. I denne sammenhengen er variansene faste størrelser. Vi slipper problemer knyttet til estimering av variansene. Effektestimatene er normalfordelte, selv for et endelig antall forsøkspersoner per studie i metaanalysene. En unngår utfordringer med asymmetri. Alle forutsetningene for å utføre en regresjonsanalyse er oppfylt. Jeg forventer således at testmetoden har et signifikansnivå som tilsvarer det nominelle. Simuleringsresultatene bekrefter dette. Nivåestimatene presenteres i Tabell 38 og Tabell 39.

I Begg og Mazumdars simuleringssituasjon kontrolleres feil av type I på et gitt nivå for både Eggers uvektede regresjonsmetode og de korrigerte rangkorrelasjonstestene. Testobservatorene kan derfor sammenliknes ved hjelp av styrken. Resultatene for de korrigerte rangkor-

relasjonstestene finnes i Tabell 20-27, Kapittel 8. Styrkeestimatene for Eggers test presenteres i Tabell 40 og Tabell 41. Selv om simuleringsscenarioet er noe kunstig, kan resultatene gi nyttig kunnskap. Verdien er størst om man vurderer resultatene i sammenheng med dem fra testsituasjonen til Macaskill et al.

Når variansspredningen er stor, vil testobservatoren til Eggers uvektede regresjonsmetode inneholde langt mer informasjon enn testobservatorene til den standardiserte og ustandardiserte testmetoden. Eksempelvis estimeres styrken til Eggers metode til 89% når $k = 25$, $\delta = 0$, variansspredningen stor og seleksjonsstyrken sterk. Til sammenlikning estimeres den tilsvarende styrken til 71% og 75 % for henholdsvis den korrigerte standardiserte og den korrigerte ustandardiserte testmetoden basert på rangkorrelasjon. Denne forskjellen avtar når variansspredningen er liten, men avhenger også av antall studier i metaanalysen og seleksjonsstyrken. Likevel er det ingen tvil om at Eggers uvektede testobservator på det jevne inneholder mest informasjon. Min tidligere anbefaling av Eggers regresjonsmetode framfor den standardiserte og ustandardiserte testen er i hovedsak begrunnet ut fra en beslutningsteoretisk analyse. En tradisjonell sammenlikning av hypotesetester, basert på styrken i simuleringssituasjonene til Begg og Mazumdar, understøtter denne konklusjonen.

Nivået estimeres også for Eggers vektede regresjonmetode i testsituasjonen til Begg og Mazumdar. Igjen velges et tosidig nominelt nivå på 0.05. Prosessen repeteres 10000 ganger. Signifikansnivået avviker stort fra det nominelle. Nivået estimeres til 0.0032 når $k = 25$, $\delta = 0$ og variansspredningen stor. Dette eksemplifiserer hvor galt det kan gå når observasjonene vektes til tross for homoskedastisitet. En rettferdig informasjonssammenlikning med de tre andre testobservatorene nevnt i denne seksjonen kan ikke gjennomføres her. Nivåestimatene støtter opp under konklusjonen fra tidligere simuleringer. Forsiktighet må vises ved bruk av Eggers vektede testmetode. Den bør generelt ikke anbefales.

Tabell 38: Nivå for Eggers uvektede regresjonsmetode. Liten metaanalyse ($k = 25$).

	Nivå [% inkluderte studier, bias]	
Varianspredning	Stor	Liten
Behandlingseffekt (δ)		
.0	4.83%	4.68%
	[100%, .00]	[100%, .00]
.5	4.94%	4.95%
	[100%, -.00]	[100%, -.00]
1.0	5.11%	4.84%
	[100%, -.00]	[100%, .00]
1.5	5.16%	4.90%
	[100%, .00]	[100%, -.00]
2.0	5.23%	5.21%
	[100%, -.00]	[100%, -.01]
2.5	4.86%	4.84%
	[100%, .00]	[100%, -.00]
3.0	4.72%	4.96%
	[100%, .00]	[100%, -.00]

Tabell 39: Nivå for Eggers uvektede regresjonsmetode. Stor metaanalyse ($k = 75$).

	Nivå [% inkluderte studier, bias]	
Varianspredning	Stor	Liten
Behandlingseffekt (δ)		
.0	5.06%	4.80%
	[100%, .00]	[100%, .00]
.5	5.07%	5.48%
	[100%, -.00]	[100%, .00]
1.0	5.18%	4.90%
	[100%, .00]	[100%, .00]
1.5	4.95%	5.22%
	[100%, .00]	[100%, -.00]
2.0	5.16%	4.89%
	[100%, -.00]	[100%, -.01]
2.5	4.94%	5.01%
	[100%, .00]	[100%, .00]
3.0	4.80%	4.99%
	[100%, .00]	[100%, -.00]

Tabell 40: Styrke for Eggers uvektede regresjonsmetode. Ensidig seleksjon basert på p-verdi. Liten metaanalyse ($k = 25$).

	Styrke [% inkluderte studier, bias]			
Seleksjonsstyrke	Sterk		Moderat	
Variansspredning	Stor	Liten	Stor	Liten
Behandlingseffekt (δ)				
.0	89%	28%	68%	17%
	[36%, .34]	[36%, .74]	[57%, .25]	[57%, .54]
.5	86%	27%	54%	15%
	[54%, .16]	[52%, .54]	[74%, .09]	[73%, .36]
1.0	74%	23%	34%	11%
	[65%, .07]	[67%, .36]	[82%, .04]	[85%, .20]
1.5	56%	17%	21%	8%
	[72%, .05]	[80%, .23]	[87%, .02]	[92%, .10]
2.0	37%	11%	13%	6%
	[78%, .03]	[88%, .14]	[90%, .02]	[96%, .05]
2.5	25%	8%	10%	5%
	[82%, .02]	[93%, .08]	[93%, .01]	[98%, .02]
3.0	18%	6%	8%	5%
	[85%, .02]	[97%, .04]	[94%, .01]	[99%, .01]

Tabell 41: Styrke for Eggers uvektede regresjonsmetode. Ensidig seleksjon basert på p-verdi. Stor metaanalyse ($k = 75$).

	Styrke [% inkluderte studier, bias]			
Seleksjonsstyrke	Sterk		Moderat	
Variansspredning	Stor	Liten	Stor	Liten
Behandlingseffekt (δ)				
.0	100%	68%	99%	45%
	[36%, .34]	[36%, .74]	[56%, .24]	[56%, .54]
.5	100%	68%	97%	39%
	[53%, .16]	[52%, .54]	[74%, .09]	[72%, .34]
1.0	100%	60%	84%	27%
	[64%, .07]	[67%, .36]	[82%, .04]	[84%, .20]
1.5	97%	45%	59%	17%
	[71%, .05]	[79%, .23]	[86%, .02]	[92%, .10]
2.0	87%	28%	39%	9%
	[77%, .03]	[88%, .13]	[90%, .02]	[96%, .05]
2.5	70%	16%	26%	6%
	[81%, .02]	[93%, .07]	[92%, .01]	[98%, .03]
3.0	54%	9%	18%	5%
	[85%, .02]	[96%, .04]	[94%, .01]	[99%, .01]

12 Ortogonal regresjon som mulig forbedring til Eggers uvektede regresjonsmetode

Vi har modellen $\mathrm{E}(Y_i|X_i = x_i) = \alpha_1 + \beta_1 x_i$, $i = 1, 2, \ldots, n$. I en metaanalysesammenheng betegner Y_i det standardiserte effektestimatet, tidligere betegnet $t_i/\sqrt{v_i}$. Variabelen X_i betegner effektestimatets presisjon, hvor jeg tidligere har nyttet notasjonen $1/\sqrt{v_i}$. I Eggers regresjonsmodell antar vi at de uavhengige variablene er målt uten feil. Med en slik antakelse er det rimelig å minimere de vertikale avstandene for å finne den "beste" linjen gjennom de observerte punktene (x_i, y_i), for $i = 1, 2, \ldots, n$. Minste kvadraters estimater for α_1 og β_1 beregnes ut fra dataene ved henholdsvis

$$\hat{\alpha}_1 = \bar{y} - \hat{\beta}_1 \bar{x}$$

og

$$\hat{\beta}_1 = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}.$$

Minste kvadraters regresjonslinje er gitt ved $\hat{y} = \hat{\alpha}_1 + \hat{\beta}_1 x$.

I en metaanalysesammenheng vil både den avhengige og den uavhengige variabelen i regresjonsmodellen være stokastiske. I testsituasjonen til Macaskill et al. [35] er responsvariabelen og forklaringsvariabelen estimert med feil. Dette kan føre til det en kaller fortynningsbias for stigningstallet og kan gi en skjevhet i estimatet for skjæringspunktet. Hvorfor nytter vi modellen $\mathrm{E}(Y|X = x) = \alpha_1 + \beta_1 x$ og ikke $\mathrm{E}(X|Y = y) = \alpha_2 + \beta_2 y$? Vi kan ikke uten videre tildele hverken X eller Y rollen som responsvariabel.

Jeg nytter den sistnevnte modellen, $\mathrm{E}(X|Y = y) = \alpha_2 + \beta_2 y$, og minimerer de horisontale avtandene. Minste kvadraters estimater for α_2 og β_2 beregnes ut fra de observerte dataene ved henholdsvis

$$\hat{\alpha}_2 = \bar{x} - \hat{\beta}_2 \bar{y}$$

og

$$\hat{\beta}_2 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(y_i - \bar{y})^2}.$$

Minste kvadraters regresjonslinje kan uttrykkes ved $\hat{x} = \hat{\alpha}_2 + \hat{\beta}_2 y$. Jeg reformulerer denne linjen slik at y er en funksjon av x. Jeg får at $\hat{y} = -\hat{\alpha}_2/\hat{\beta}_2 + x/\hat{\beta}_2$. De to regresjonslinjene er like hvis og bare hvis $\hat{\alpha}_1 = -\hat{\alpha}_2/\hat{\beta}_2$ og $\hat{\beta}_1 = \hat{\beta}_2^{-1}$. Dette kan ikke antas generelt.

Det er forstyrrende at de to minste kvadraters regresjonslinjene er ulike når en ikke uten videre kan anta at Y er responsvariabelen og X forklaringsvariabelen. Ortogonal minste kvadraters metode, også kjent som ortogonal regresjon, kan være en mulig løsning på problemet. Vi finner linjen som minimerer de ortogonale avstandene til de observerte punktene (x_i, y_i), $i = 1, 2, \ldots, n$.

12.1 Ortogonal minste kvadraters metode

Jeg tar utgangspunkt i utledningen gitt av Casella og Berger [10]. Forfatterne overlater enkelte mellomregninger til oppgaveseksjonen. Jeg løser disse oppgavene og inkluderer løsningene i utledningen nedenfor.

La $(\hat{x}', \hat{y}')$ være den ortogonale projeksjonen av punktet (x', y') på linjen $l : y = \alpha + \beta x$. Videre ligger (x', y') på en linje $m : y = \alpha_m + \beta_m x$. Linjene m og l antas ortogonale. Stigningstallet til linjen m finnes ved å løse likningen $\beta\beta_m = -1$, slik at $\beta_m = -\beta^{-1}$. Fordi (x', y') er et punkt på m, vet vi at likningen $y' = \alpha_m - x'/\beta$ må være oppfylt. Dette gir at $\alpha_m = y' + x'/\beta$. Linjen m kan da uttrykkes ved $m : y = y' + x'/\beta - x/\beta$.

Koordinatene til den ortogonale projeksjonen av (x', y') på l finnes ved å løse likningssystemet

$$\begin{aligned} y &= y' + \frac{x'}{\beta} - \frac{x}{\beta} \\ y &= \alpha + \beta x. \end{aligned}$$

Enkel algebra gir

$$\hat{x}' = \frac{\beta y' + x' - \alpha\beta}{1 + \beta^2} \tag{5}$$

og

$$\hat{y}' = \alpha + \frac{\beta}{1 + \beta^2}(\beta y' + x' - \alpha\beta). \tag{6}$$

Vi observerer $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Den kvadratiske avstanden fra et observert punkt (x_i, y_i) og linjen $y = \alpha + \beta x$ er gitt ved $(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2$, hvor $\hat{x}_i$ og $\hat{y}_i$ defineres ved uttrykkene (5) og (6) henholdsvis. Jeg finner de verdiene av α og β som minimerer

$$S = \sum_{i=1}^{n} \big((x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2\big).$$

Ved innsetting av uttrykket for $\hat{x}_i$ og $\hat{y}_i$ får jeg at

$$\begin{aligned} S &= \sum \left(x_i - \frac{\beta y_i + x_i - \alpha\beta}{1 + \beta^2} \right)^2 + \sum \left(y_i - \alpha - \frac{\beta}{1 + \beta^2}(\beta y_i + x_i - \alpha\beta) \right)^2 \\ &= \frac{1}{1 + \beta^2} \sum (y_i - \alpha - \beta x_i)^2. \end{aligned}$$

Mellomregninger er utelatt.

Jeg partiellderiverer S med hensyn på α og får at

$$\frac{\partial S}{\partial \alpha} = -\frac{2}{1 + \beta^2} \sum (y_i - \alpha - \beta x_i).$$

Dette uttrykket settes lik null og forenkles. Minste kvadraters estimat for α kan beregnes ved $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$.

Ved innsetting av $\alpha = \bar{y} - \beta\bar{x}$ i uttrykket for S får jeg at

$$\begin{aligned} S &= \frac{1}{1+\beta^2} \sum \big((y_i - \bar{y}) - \beta(x_i - \bar{x})\big)^2 \\ &= \frac{1}{1+\beta^2} \sum \big((y_i - \bar{y})^2 - 2\beta(x_i - \bar{x})(y_i - \bar{y}) + \beta^2(x_i - \bar{x})^2\big). \end{aligned}$$

Jeg partiellderiverer S med hensyn på β, slik at

$$\begin{aligned} \frac{\partial S}{\partial \beta} = &-\frac{2\beta}{(1+\beta^2)^2} \Big(\sum (y_i - \bar{y})^2 - 2\beta \sum (x_i - \bar{x})(y_i - \bar{y}) + \beta^2 \sum (x_i - \bar{x})^2\Big) \\ &+ \frac{1}{1+\beta^2} \Big(-2 \sum (x_i - \bar{x})(y_i - \bar{y}) + 2\beta \sum (x_i - \bar{x})^2\Big). \end{aligned}$$

Uttrykket settes lik null. Etter noe algebra står jeg igjen med

$$\beta^2 \sum (x_i - \bar{x})(y_i - \bar{y}) + \beta \Big(\sum (x_i - \bar{x})^2 - \sum (y_i - \bar{y})^2\Big) - \sum (x_i - \bar{x})(y_i - \bar{y}) = 0,$$

et andregradsuttrykk for β. Jeg definerer

$$S_{yy} = \sum (y_i - \bar{y})^2,\ S_{xx} = \sum (x_i - \bar{x})^2 \text{ og } S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}).$$

Jeg løser andregradslikningen for β og får at

$$\hat{\beta} = \frac{S_{yy} - S_{xx} \pm \sqrt{(S_{yy} - S_{xx})^2 + 4S_{xy}^2}}{2S_{xy}}.$$

Vi har to kritiske punkt for S. Det ene er et minimumspunkt, det andre et maksimum. Jeg ønsker å finne den verdien av β som gir et minimum. Jeg legger merke til at $S \to S_{xx}$ når $\beta \to \infty$ og $\beta \to -\infty$. I tillegg er $S = S_{xx} - \big(\sqrt{(S_{yy} - S_{xx})^2 + 4S_{xy}^2}\big)/(1+\beta^2) < S_{xx}$ når $\beta = \big(S_{yy} - S_{xx} + \sqrt{(S_{yy} - S_{xx})^2 + 4S_{xy}^2}\big)/(2S_{xy})$. Når $\beta = \big(S_{yy} - S_{xx} - \sqrt{(S_{yy} - S_{xx})^2 + 4S_{xy}^2}\big)/(2S_{xy})$, har vi at $S = S_{xx} + \big(\sqrt{(S_{yy} - S_{xx})^2 + 4S_{xy}^2}\big)/(1+\beta^2) > S_{xx}$. Det følger at minste kvadraters estimat for β kan beregnes ut fra de observerte dataene ved

$$\hat{\beta} = \frac{S_{yy} - S_{xx} + \sqrt{(S_{yy} - S_{xx})^2 + 4S_{xy}^2}}{2S_{xy}}.$$

12.2 Testobservator for inferens om skjæringspunktet, samt dens fordeling

I likhet med Eggers regresjonsmetode er jeg interessert i å teste nullhypotesen om at α er lik null. Jeg må finne en testobservator til dette formålet. En utledning er omfattende, og jeg presenterer i stedet asymptotiske resultater fra Fuller [18]. Mange nye størrelser defineres, men poenget er bare å referere Fullers resultat.

La $Y_i = \alpha + \beta X_i + \epsilon_i$ og $X_i' = X_i + e_i$. I denne seksjonen er det X_i' som observeres, mens X_i er den sanne verdien av forklaringsvariabelen. Jeg definerer $\boldsymbol{\delta}_i = (\epsilon_i, X_i, e_i)^T$. De stokastiske variablene, ϵ_i, X_i og e_i, er uavhengige, og $\boldsymbol{\delta}_i$ har en simultan multivariat normalfordeling. Forventningen til $\boldsymbol{\delta}_i$ er gitt ved $\boldsymbol{\mu} = (0, \mu_X, 0)^T$, og kovariansmatrisen defineres ved $\boldsymbol{\Sigma} = \text{diag}(\sigma_{\epsilon\epsilon}, \sigma_{XX}, \sigma_{ee})$, $i = 1, 2, \ldots, n$. Fordi X_i er stokastisk, har vi det som kalles en strukturell modell. Denne modellen står i kontrast til den funksjonelle modellen, hvor vi regner betinget gitt $X_i = x_i$. Vi antar at $\sigma_{\epsilon\epsilon} = \sigma_{ee}$ og selvfølgelig at $\sigma_{\epsilon\epsilon}, \sigma_{XX} > 0$. Vi definerer $\boldsymbol{\theta} = (\alpha, \beta)^T$ og $\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\beta})^T$, hvor $\hat{\alpha}$ og $\hat{\beta}$ i denne seksjonen er definert ved henholdsvis

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}'$$

og

$$\hat{\beta} = \frac{S_{YY} - S_{X'X'} + \sqrt{(S_{YY} - S_{X'X'})^2 + 4S_{X'Y}^2}}{2S_{X'Y}}.$$

Fra nå av defineres $S_{X'Y} = \sum(X_i' - \bar{X}')(Y_i - \bar{Y})/(n-1)$, $S_{X'X'} = \sum(X_i' - \bar{X}')^2/(n-1)$ og $S_{YY} = \sum(Y_i - \bar{Y})^2/(n-1)$.

Når $n \to \infty$, konvergerer $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ i fordeling mot $\text{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\theta\theta})$, hvor

$$\boldsymbol{\Sigma}_{\theta\theta} = \begin{pmatrix} \sigma_{vv} + \mu_X^2\sigma_{\beta\beta} & -\mu_X\sigma_{\beta\beta} \\ -\mu_X\sigma_{\beta\beta} & \sigma_{\beta\beta} \end{pmatrix}.$$

Vi definerer $v_i = \epsilon_i - \beta e_i$, $\sigma_{vv} = \sigma_{\epsilon\epsilon} + \beta^2\sigma_{ee}$, $\sigma_{ev} = -\beta\sigma_{ee}$ og $\sigma_{\beta\beta} = \sigma_{XX}^{-2}(\sigma_{XX}\sigma_{vv} + \sigma_{ee}\sigma_{vv} - \sigma_{ev}^2)$.

En estimator for $\boldsymbol{\Sigma}_{\theta\theta}$ er gitt ved

$$\hat{\boldsymbol{\Sigma}}_{\theta\theta} = \begin{pmatrix} n^{-1}m_{vv} + \bar{X}'^2\hat{\sigma}_{\hat{\beta}\hat{\beta}} & -\bar{X}'\hat{\sigma}_{\hat{\beta}\hat{\beta}} \\ -\bar{X}'\hat{\sigma}_{\hat{\beta}\hat{\beta}} & \hat{\sigma}_{\hat{\beta}\hat{\beta}} \end{pmatrix}.$$

Her er $\hat{\sigma}_{\hat{\beta}\hat{\beta}} = (n-1)^{-1}\hat{\sigma}_{XX}^{-2}(\hat{\sigma}_{XX}m_{vv} + \hat{\sigma}_{ee}m_{vv} - \hat{\sigma}_{ev}^2)$, $m_{vv} = (n-2)^{-1}(n-1)(1+\hat{\beta}^2)\hat{\sigma}_{ee}$ og $\hat{\sigma}_{ev} = -\hat{\beta}\hat{\sigma}_{ee}$. Videre defineres $\hat{\sigma}_{XX} = 2^{-1}\big(((S_{YY} - S_{X'X'})^2 + 4S_{X'Y}^2)^{1/2} - (S_{YY} - S_{X'X'})\big)$ og $\hat{\sigma}_{ee} = 2^{-1}\big(S_{YY} + S_{X'X'} - ((S_{YY} - S_{X'X'})^2 + 4S_{X'Y}^2)^{1/2}\big)$.

Ifølge Fuller [18] er

$$t = \frac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}_{\hat{\beta}\hat{\beta}}}}$$

tilnærmet standardnormalfordelt. For liten n foreslår forfatteren å tilnærme fordelingen til t ved hjelp av en t-fordeling med $n-2$ frihetsgrader. I tråd med teorien gitt hos Fuller vil også

$$t = \frac{\hat{\alpha} - \alpha}{\sqrt{\hat{\sigma}_{\hat{\alpha}\hat{\alpha}}}}$$

være tilnærmet standardnormalfordelt, hvor $\hat{\sigma}_{\hat{\alpha}\hat{\alpha}} = n^{-1}m_{vv} + \bar{X}'^2\hat{\sigma}_{\hat{\beta}\hat{\beta}}$. Videre kan en, for liten n, tilnærme fordelingen til t ved en t-fordeling med $n-2$ frihetsgrader. Dessverre vet

vi ikke hvor god tilnærming en t-fordeling vil være. Forsiktighet bør vises. Jeg mistenker trykkfeil ved beregning av denne testobservatoren i eksempel 1.3.2 hos Fuller [18]. Fuller tester hypotesen $H_0 : \alpha = 0$. Han har tidligere estimert $\hat{\alpha} = -4.1686$ og $\hat{\sigma}_{\hat{\alpha}\hat{\alpha}} = 1.2720$, men beregner $t = (1.2608)^{-1/2}(-4.1686)$. Det står ikke forklart hvordan Fuller finner verdien 1.2608.

Prinsipalkomponentanalyse samsvarer med ortogonal regresjon. I to dimensjoner korresponderer stigningstallet til den ortogonale regresjonslinjen med stigningstallet til den første prinsipale komponenten. Eksisterende resultater fra inferens om den første prinsipale komponenten kan brukes til å utføre inferens om β [30]. Jeg vil ikke utdype denne tematikken videre.

12.3 Kommentarer til bruk av ortogonal regresjon ved testing for publikasjonsbias

Det gir intuitivt mening å forsøke å forbedre Eggers regresjonsmetode ved bruk av ortogonal regresjon. Både den uavhengige og den avhengige variabelen i regresjonsmodellen inneholder målefeil. For å utføre inferens om α i en ortogonal regresjonsanalyse, er det nødvendig med flere antakelser. Blant annet er det et krav om at X_i, ϵ_i og e_i, definert i forrige seksjon, skal være uavhengige stokastiske variable. Det er rimelig å anta at X_i, nå effektestimatets virkelige presisjon, er en stokastisk variabel i reelle metaanalyser. For 2×2-tabeller er den uavhengige variabelen i regresjonsmodellen et biased estimat av den virkelige presisjonen. Det følger at X_i og e_i ikke er stokastisk uavhengige. Jeg har tidligere diskutert at målefeil ved estimering av effekten kan forplante seg videre ved estimering av variansen. Av den grunn kan jeg heller ikke anta uavhengighet mellom ϵ_i og e_i. Det er vanskelig å forutsi hvordan brudd på forutsetningene vil påvirke simuleringsresultatene. I tillegg må en ha i bakhodet at asymmetriske funnelplott ikke nødvendigvis er et optimalt kriterium for å identifisere selektiv publikasjon, jamfør Seksjon 11.3.

Ortogonal minste kvadraters metode er ikke den eneste metoden som kan brukes til å løse problemet med målefeil i den uavhengige variabelen. Eksempelvis kan en også nytte maksimum-likelihood estimering eller momentmetoden. Casella og Berger [10] utleder maksimum-likelihood-estimatorer, mens Fuller [18] utleder estimatorer ved hjelp av momentmetoden. I dette tilfellet vil estimatorene utledet ved de to ulike metodene sammenfalle. Disse estimeringsmetodene antar $\sigma_{\epsilon\epsilon} = \lambda\sigma_{ee}$, hvor $\lambda > 0$ er en kjent konstant. Dersom $\sigma_{\epsilon\epsilon} = \sigma_{ee}$, vil maksimum likelihood og momentmetoden gi de samme estimatorene for α og β som ortogonal minste kvadraters metode.

I flere situasjoner vil $\lambda = 1$ være en urimelig antakelse. En kan derfor argumentere for at den ortogonale regresjonsmetoden er noe for enkel. Fordi jeg allerede har beveget meg bort fra en korrekt sannsynlighetsmodell, gir det likevel mening å forsøke med det enkleste alternativet.

Jeg unngår ytterlige utfordringer knyttet til estimering av λ.

12.4 Simuleringsresultater og vurdering av testmetoden

Den ortogonale testmetoden vurderes i simuleringsscenarioet til Macaskill et al [35], ved konfigurasjon A og B. Simuleringsprosedyren repeteres 10000 ganger. Jeg velger stadig et nominelt nivå på 0.05 ved utføring av ensidige tester og et nominelt nivå på 0.10 ved utføring av tosidige tester. Som vanlig avgrenses simuleringene til ensidig seleksjon basert på p-verdi. Simuleringene utføres i R [43], hvor jeg nytter formlene utledet i dette kapittelet.

Tabell 42 viser de gjennomsnittlige skjæringspunktene for den ortogonale regresjonsmetoden, i tilfeller hvor metaanalysene ikke er utsatt for selektiv publikasjon. Resultatene understøtter at estimatoren ikke er forventningsrett når $\delta \neq 0$. Dessuten antyder store maksimale standardavvik at metoden er upålitelig. Når den underliggende effekten er ulik null, er de gjennomsnittlige skjæringspunktene positive. Det er ikke overraskende om vi, under nullhypotesen om ingen publikasjonsbias, forkaster flere metaanalyser under halen som representerer $\alpha > 0$ enn halen som representerer $\alpha < 0$.

Nivåestimatene er presentert i Tabell 43. Estimatene underbygger antakelsene i avsnittet over. Testmetoden ser ut til å være uforutsigbar. Den ortogonale regresjonsmodellen er ikke tilpasset den aktuelle situasjonen.

Tabell 44 viser styrkeestimatene for konfigurasjon A og B i testsituasjonen gitt av Macaskill et al. [35]. Biasen og andelen av de genererte studiene som inkluderes i metaanalysene tilsvarer resultatene gitt i Tabell 31. Estimatene refereres ikke her.

Jeg vurderer først den ensidige testen når det ikke er noen underliggende effekt ved konfigurasjon A. Her forkastes ikke nullhypotesen om at det virkelige nivået er lik det nominelle ved et tosidig signifikansnivå på 0.05. I dette tilfellet ser vi at Eggers uvektede testobservator inneholder mer informasjon enn den ortogonale testobservatoren. Når $\delta \neq 0$, vil den ortogonale testens virkelige ensidige nivå være langt lavere enn det nominelle, både ved konfigurasjon A og B. Det er ikke uventet at den ensidige ortogonale testen har lav styrke.

Den tosidige styrken er dårligere sammenliknet med Eggers uvektede regresjonsmetode når $\delta = \ln(1), \ln(2/3), \ln(1/2)$. Når $\delta = \ln(1/4)$, vil den ortogonale regresjonsmetoden ha bedre tosidig styrke. Ut fra de ensidige testene, ser vi at dette skyldes forkastning i "feil" hale. Her blir vi atter en gang påminnet viktigheten av å vurdere styrken i sammenheng med nivået.

Ortogonal regresjon kan ikke anbefales som basis for en testmetode som tester for publikasjonsbias i metaanalyser. Jeg ser ikke behov for en mer inngående diskusjon rundt simuleringsresultatene, ei heller nytteverdien av å presentere simuleringsresultater for de resterende konfigurasjonene.

Tabell 42: Gjennomsnittlig skjæringspunkt for testmetoden basert på ortogonal regresjon.

Behandlingseffekt (δ)	Konfigurasjon	Gjennomsnittlig skjæringspunkt
ln(1)		
	A	-2.551
	B	-8.0×10^{-3}
ln(2/3)		
	A	3.909^{*}
	B	$5.3 \times 10^{-1*}$
ln(1/2)		
	A	2.157^{*}
	B	$6.4 \times 10^{-1*}$
ln(1/4)		
	A	1.354^{*}
	B	$5.8 \times 10^{-1*}$
Maksimal SE		
	A	2.048
	B	8.3×10^{-3}

Maksimal SE = maksimal standardfeil for parameterestimatene.
* Estimatorene inneholder statistisk signifikant bias ($|z| = |\text{gjennomsnitt}|/\text{SE} > z_{\alpha/2} = 1.96$).
Konfigurasjon A (21 studier: 11 × 100/gruppe, 6 × 200/gruppe, 4 × 300/gruppe).
Konfigurasjon B (21 studier: 10 × 100/gruppe, 5 × 200/gruppe, 3 × 300/ gruppe, 2 × 500/ gruppe, 1 × 1000/gruppe).

Tabell 43: Nivåestimat for testmetoden basert på ortogonal regresjon.

Behandlingseffekt (δ)	Konfigurasjon	Nivå % Ensidig	Nivå % Andre hale
ln(1)			
	A	**5.3**	6.0
	B	**6.5**	5.7
ln(2/3)			
	A	**0.8**	19.8
	B	**1.8**	15.5
ln(1/2)			
	A	**0.4**	28.7
	B	**0.9**	20.5
ln(1/4)			
	A	**0.4**	27.9
	B	**1.3**	20.8

Konfigurasjon A (21 studier: 11 × 100/gruppe, 6 × 200/gruppe, 4 × 300/gruppe).
Konfigurasjon B (21 studier: 10 × 100/gruppe, 5 × 200/gruppe, 3 × 300/ gruppe, 2 × 500/ gruppe, 1 × 1000/gruppe).

Tabell 44: Styrkeestimat for testmetoden basert på ortogonal regresjon.

Behandlingseffekt (δ)	Konfigurasjon	Styrke % Ensidig	Styrke % Andre hale
ln(1)			
	A	**21.9**	0.9
	B	**48.8**	0.1
ln(2/3)			
	A	**5.5**	4.5
	B	**14.3**	1.8
ln(1/2)			
	A	**1.0**	13.5
	B	**2.9**	9.0
ln(1/4)			
	A	**0.5**	23.5
	B	**1.4**	18.3

Konfigurasjon A (21 studier: 11 × 100/gruppe, 6 × 200/gruppe, 4 × 300/gruppe).
Konfigurasjon B (21 studier: 10 × 100/gruppe, 5 × 200/gruppe, 3 × 300/ gruppe, 2 × 500/ gruppe, 1 × 1000/gruppe).

13 Oppsummering og videre arbeid

Forsiktighet bør vises ved bruk av Begg og Mazumdars rangkorrelasjonstest når en tester for publikasjonsbias i metaanalyser. Simuleringsresultater viser at testens virkelige signifikansnivå generelt ikke er lik det nominelle. Nivået er lavere enn det nominelle i testsituasjonen til Begg og Mazumdar. Matematisk har jeg vist at dette skyldes den betingede korrelasjonen mellom de standardiserte effektestimatene gitt variansene,

$$\mathrm{Cor}(t_i^*, t_j^* | v_1, ..., v_k) = -\frac{1}{(v_i^*)^{1/2}(v_j^*)^{1/2} \sum v_l^{-1}}.$$

Konsekvensene av testmetodens ulemper er ikke neglisjerbare. En bør ta hensyn til de noe uvanlige rollene til nullhypotesen og den alternative hypotesen. Den generelle litteraturen [5, 8, 35, 41] ser ut til å glemme dette.

Forbedringer til Begg og Mazumdars rangkorrelasjonstest er foreslått i situasjoner hvor effektestimatene kan antas å være normalfordelte under nullhypotesen om ingen publikasjonsbias. Jeg anbefaler å tilpasse nivået ved hjelp av den simulerte fordelingen til Kendalls tau, betinget på de estimerte variansene. Om en ikke korrigerer nivået, anbefales den ustandardiserte testen. Her korreleres effektestimatene mot de tilhørende variansene. Effektestimatene standardiseres ikke. Simuleringsresultater viser at den ustandardiserte testen også kan foretrekkes framfor metoden introdusert av Begg og Mazumdar i tilfeller hvor effektestimatene ikke nødvendigvis er symmetrisk fordelte under nullhypotesen. Anbefalingen av den ustandardiserte testmodellen er basert på en hypotetisk og uformell beslutningsteoretisk analyse.

Simuleringsresultatene avhenger av de ulike simuleringsscenarioene. Det er behov for videre testing av de korrigerte rangkorrelasjonstestene i andre og mer reelle metaanalysesituasjoner enn dem jeg har studert. Det er særlig behov for videre testing i tilfeller hvor effektestimatene ikke kan antas å være normalfordelte under nullhypotesen om ingen publikasjonsbias. Vil mine konklusjoner gjelde på et mer generelt grunnlag? Kan vi utbedre de korrigerte testmetodene?

Jeg foreslår en regresjonsbasert testmetode som er en forbedring til regresjonsmetodene introdusert av Macaskill et al. Fordelen med enkelte av metodene basert på funnelplottregresjon er at en i større grad klarer å kontrollere feil av type I på et gitt nivå. Nivået er tilnærmet korrekt, selv når effektestimatene ikke har en symmetrisk fordeling under nullhypotesen om ingen publikasjonsbias.

Det er kun i enkelte tilfeller vi kan gjennomføre en rettferdig sammenlikning av informasjonen til de ulike testobservatorene introdusert i oppgaven. I disse situasjonene er det Eggers uvektede testobservator som inneholder mest informasjon. Det er ønskelig å arbeide videre med denne metoden og forsøke å tilpasse nivået. I testsituasjonen til Macaskill et al. avviker det virkelige nivået til Eggers testmetode mer fra det nominelle desto lenger bort fra

nullverdien den underliggende, sanne effekten ligger. I disse tilfellene vil funnelplottet være asymmetrisk, selv uten publikasjonsbias. Er asymmetri i funnelplottet et passende kriterium for identifisering av publikasjonsbias i metaanalyser?

Samtlige testmetoder i denne oppgaven kritiseres for lav styrke. Lav styrke er særlig en utfordring i tilfeller hvor sampelstørrelsen er liten. Jeg har gjentatte ganger poengtert viktigheten av å tolke resultatene med forsiktighet. En bør ikke utelukke publikasjonsbias, selv om nullhypotesen om ingen publikasjonsbias ikke forkastes. Testprosedyrene er likevel viktige, objektive supplement til en visuell inspeksjon av funnelplottet for å identifisere publikasjonsbias i metaanalyser [35].

Metoder for fellesestimering som forsøker å korrigere for selektiv publikasjon er et sentralt tema innen publikasjonsbias i metaanalyser. Duval og Tweedie [13, 14] introduserer en metode for fellesestimering som er basert på symmetriegenskapene til funnelplottet. Denne metoden kan muligens være et nyttig tillegg til testmetodene introdusert i denne oppgaven. Egenskapene til trim-and-fill-metoden bør derimot undersøkes i tilfeller hvor en ikke har symmetriske funnelplott under nullhypotesen om ingen publikasjonsbias.

Jeg vil bemerke at publikasjonsbias ikke er den eneste faktoren som kan medføre asymmetriske funnelplott. Statistisk heterogenitet eksisterer når den sanne effekten som blir evaluert er ulik i forskjellige studier [21]. Denne faktoren kan også gi funnelplott som ikke er symmetriske [15].

Intuitivt har en grunn til å frykte at studier med liten sampelstørrelse kan ha større problemer med å korrigere for ulikheter innad i studien enn studier med stor sampelstørrelse. Disse ulikhetene kan medføre ekstra variasjon i effektestimatet til små studier, og også ekstra (klinisk) heterogenitet mellom de ulike studiene i en metaanalyse. Jeg har en idé om å teste ut robust regresjon [23] som et alternativ til de eksisterende regresjonsmetodene for å identifisere publikasjonsbias i metaanalyser hvor denne tilleggsheterogeniteten eksisterer. Robust regresjon kan utføres med utgangspunkt i alle de regresjonsbaserte testmetodene introdusert i Kapittel 10. Ekstra variasjon mellom studiene vil ikke oppstå i metaanalysesituasjonene jeg har studert i denne oppgaven. Simuleringsscenarioene er i dette henseende noe forenklet sammenliknet med reelle metaanalyser. Robust regresjon, som basis for en modell som tester for publikasjonsbias i metaanalyser, må derfor undersøkes ved en senere anledning.

A Kort om konfidensintervaller

I denne oppgaven er det behov for å teste hypoteser på bakgrunn av simulerte data. Eksempelvis ønsker jeg å teste nullhypotesen om at det virkelige signifikansnivået er lik det nominelle. Det er en nøye sammenheng mellom hypotesetester og konfidensintervaller.

Lillestøl [32] definerer et konfidensintervall for en ukjent parameter θ som et intervall med grenser, som med en gitt sannsynlighet c, omslutter θ. Sannsynligheten c kalles konfidensnivået. Et konfidensintervall for θ med konfidensnivå $100(1-\alpha)$ kan nyttes direkte til å ta stilling til en hypotese med signifikansnivå α. Konfidensintervallet sier noe om hvilke verdier av θ en tror på.

Nullhypotesen, $H_0 : \theta = \theta_0$, skal testes mot den alternative hypotesen, $H_1 : \theta \neq \theta_0$. Da kan en konstruere et tosidig konfidensintervall for θ med konfidensnivå $100(1-\alpha)$. Dersom θ_0 befinner seg innenfor konfidensintervallets grenser, kan en ikke utelukke at θ_0 er den sanne verdien. Vi forkaster ikke nullhypotesen med signifikansnivå α. Hvis derimot θ_0 ikke er å finne innenfor konfidensintervallets grenser, er vi $100(1-\alpha)\%$ sikre på at θ_0 ikke er den sanne verdien. Vi forkaster nullhypotesen med signifikansnivå α. Dersom vi formulerer en ensidig alternativ hypotese, tar vi utgangspunkt i et ensidig konfidensintervall.

Vi observerer en binomisk forsøksrekke med n uavhengige forsøk og registrerer Y lik antall ganger hendelsen inntreffer. Sannsynligheten for suksess i hvert forsøk er p. Den stokastiske variabelen Y er binomisk fordelt. Vi ønsker å bestemme presisjonen til den relative frekvensen Y/n som en estimator for p. Dette kan løses ved å konstruere et konfidensintervall for den ukjente parameteren p, basert på Y/n.

Sannsynlighetsfordelingen til en binomisk variabel Y er

$$P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y = 0, 1, \ldots, n, \quad 0 \leq p \leq 1.$$

Videre er $\mathrm{E}(Y) = np$ og $\mathrm{Var}(Y) = np(1-p)$. Det klassiske sentralgrenseteoremet gir at $\big((Y/n) - p\big)/\big(\sqrt{p(1-p)/n}\big)$ konvergerer mot standardnormalfordelingen når $n \to \infty$.

Siden $\big((Y/n) - p\big)/\big(\sqrt{p(1-p)/n}\big)$ er asymptotisk normalfordelt, kan vi, for en gitt sannsynlighet $1-\alpha$, finne en $z_{\alpha/2}$ slik at

$$P\left(-z_{\alpha/2} \leq \frac{(Y/n) - p}{\sqrt{p(1-p)/n}} \leq z_{\alpha/2}\right) \approx 1 - \alpha.$$

Her er z_α et tall slik at $P(Z \geq z_\alpha) = \alpha$, hvor Z er standardnormalfordelt. Enkel algebra gir

$$P\left(\frac{Y}{n} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \frac{Y}{n} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) \approx 1 - \alpha.$$

Parameteren p er ukjent. Vi kan derfor ikke bruke p til å beregne endepunktene. Det er flere måter å gå fram på for å løse dette problemet. Jeg velger å approksimere p i endepunktene med Y/n. Da får vi

$$P\left(\frac{Y}{n} - z_{\alpha/2}\sqrt{\frac{(Y/n)\big(1-(Y/n)\big)}{n}} \le p \le \frac{Y}{n} + z_{\alpha/2}\sqrt{\frac{(Y/n)\big(1-(Y/n)\big)}{n}}\right) \approx 1-\alpha.$$

For stor n er da

$$\left[\frac{y}{n} - z_{\alpha/2}\sqrt{\frac{(y/n)\big(1-(y/n)\big)}{n}}\ ,\ \frac{y}{n} + z_{\alpha/2}\sqrt{\frac{(y/n)\big(1-(y/n)\big)}{n}}\right]$$

et tilnærmet $100(1-\alpha)\%$ konfidensintervall for p, hvor y er den observerte verdien av Y. Heltallskorreksjon kan forbedre normaltilnærmingen.

Ensidige konfidensintervall for p er approksimert ved

$$\left[0\ ,\ \frac{y}{n} + z_{\alpha}\sqrt{\frac{(y/n)\big(1-(y/n)\big)}{n}}\right]$$

og

$$\left[\frac{y}{n} - z_{\alpha}\sqrt{\frac{(y/n)\big(1-(y/n)\big)}{n}}\ ,\ 1\right].$$

Det første intervallet gir en øvre grense for p, det siste gir en nedre grense for p.

Jeg presiserer at konfidensintervallene er tilnærmede. For det første nytter jeg at $\big((Y/n)-p\big)\big(\sqrt{p(1-p)/n}\big)$ er asymptotisk normalfordelt. For det andre approksimerer jeg p i endepunktene. Vollset [55] argumenterer for at dette intervallet, samt dets versjon med heltallskorrigering, ikke bør brukes. Hvis antall suksesser Y er lik 0 eller lik antall forsøk, n, vil disse metodene ikke produsere noe intervall. Dessuten vil konfidensintervallene ikke være troverdige om p er nær 0 eller 1, samtidig som antall suksesser har en tendens til å ligge nær 0 eller n. Vollset [55] viser at normaltilnærmingsmetoden ikke fungerer bra for små n. Derimot vil metoden fungere langt bedre når n er stor, bortsett fra når estimatoren for p er nær endepunktene. Dette er ikke uventede resultater ut fra de tilnærmingene som er gjort i utledningen av konfidensintervallene.

Pires og Amado [42] argumenterer på sin side at små modifikasjoner i grenseverdiene, samtidig med en kontinuitetskorrekasjon, fører til at normaltilnærmingsintervallene vil gi akseptable resultater. Vollset [55] og Pires og Amado [42] foreslår forbedringer til normaltilnærmingsintervallene. Jeg velger likevel å holde meg til den enkle lærebokmetoden. Jeg arbeider med store utvalg, samtidig som antall suksesser sjelden vil ligge nær endepunktene.

Hva om det finnes flere uavhengige metoder å gjennomføre et eksperiment på, og vi ønsker å sammenlikne to av disse metodene? Da kan vi konstruere et tosidig konfidensintervall for $p_1 - p_2$, hvor p_1 er sannsynligheten for suksess for den første metoden, og p_2 er sannsynligheten for suksess for den andre metoden. Nullhypotesen om at $p_1 = p_2$ forkastes dersom konfidensintervallet ikke inneholder tallet 0.

Framgangsmåten for å finne dette intervallet tilsvarer utledningen av konfidensintervallet for p. Antall uavhengige forsøk for den første og andre metoden beskrives ved henholdsvis n_1 og n_2. Disse forsøkene resulterer i Y_1 og Y_2 suksesser. Forventningen og variansen til Y_1/n er p_1 og $p_1(1-p_1)/n_1$ henholdsvis. Tilsvarende er $\mathrm{E}(Y_2/n) = p_2$ og $\mathrm{Var}(Y_2/n) = p_2(1-p_2)/n_2$. Da får vi at

$$\mathrm{E}\big((Y_1/n) - (Y_2/n)\big) = \mathrm{E}(Y_1/n_1) - \mathrm{E}(Y_2/n_2) = p_1 - p_2.$$

Videre er

$$\mathrm{Var}\big((Y_1/n_1) - (Y_2/n_2)\big) = \mathrm{Var}(Y_1/n_1) + \mathrm{Var}(Y_2/n_2) = \big(p_1(1-p_1)/n_1\big) + \big(p_2(1-p_2)/n_2\big)$$

siden Y_1 og Y_2 er uavhengige.

Både Y_1/n_1 og Y_2/n_2 er asymptotisk normalfordelte. En lineær kombinasjon av uavhengige, asymptotiske normalfordelte variabler er også asymptotisk normalfordelt. Det følger at $\big((Y_1/n_1) - (Y_2/n_2) - (p_1 - p_2)\big)/\big(\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}\big)$ er tilnærmet standardnormalfordelt når n_1 og n_2 er store. Av samme grunn som tidligere erstatter vi p_1 og p_2 i nevneren med henholdsvis Y_1/n_1 og Y_2/n_2. For en gitt sannsynlighet $1-\alpha$ kan vi finne en $z_{\alpha/2}$ slik at

$$P\left(-z_{\alpha/2} \le \frac{(Y_1/n_1) - (Y_2/n_2) - (p_1 - p_2)}{\sqrt{(Y_1/n_1)\big(1-(Y_1/n_1)\big)/n_1 + (Y_2/n_2)\big(1-(Y_2/n_2)\big)/n_2}} \le z_{\alpha/2}\right) \approx 1-\alpha.$$

Den observerte verdien av Y_1 er y_1, mens y_2 er den observerte verdien av Y_2. Et konfidensintervall for $p_1 - p_2$ med konfidensnivå $100(1-\alpha)$ finnes ved hjelp av enkel algebra og kan uttrykkes ved

$$\left[\frac{y_1}{n_1} - \frac{y_2}{n_2} - z_{\alpha/2}\sqrt{\frac{(y_1/n_1)\big(1-(y_1/n_1)\big)}{n_1} + \frac{(y_2/n_2)\big(1-(y_2/n_2)\big)}{n_2}}\ ,\right.$$
$$\left.\frac{y_1}{n_1} - \frac{y_2}{n_2} + z_{\alpha/2}\sqrt{\frac{(y_1/n_1)\big(1-(y_1/n_1)\big)}{n_1} + \frac{(y_2/n_2)\big(1-(y_2/n_2)\big)}{n_2}}\ \right].$$

Ved noen anledninger i oppgaven konstrueres konfidensintervaller for den ukjente forventningen og den ukjente variansen. Jeg vil kort gi uttrykkene for disse intervallene, men ser ikke behov for en inngående beskrivelse.

Et tosidig konfidensintervall for den ukjente forventningen, μ, med konfidensnivå $100(1-\alpha)$ er gitt ved $\bar{x} \pm z_{\alpha/2}\sigma/\sqrt{n}$. Her antas $X_1, X_2, \ldots, X_n$ å være iid variable fra en normalfordeling

med forventning μ og varians σ^2. Estimatoren $\bar{X}$ er forventningsrett for μ. Variansen er kjent, og $\bar{x}$ er den obsertverte verdien av $\bar{X}$.

Den underliggende fordelingen er ikke alltid normal. Grunnet Sentralgrenseteoremet vil konfidensintervallet, $\bar{x} \pm z_{\alpha/2}\sigma/\sqrt{n}$, likevel fungere som en grei tilnærming dersom n er stor. Det samme er tilfellet om variansen er ukjent og n stor. Da må derimot σ ersattes med s, hvor s er den observerte verdien av sampelstandardavviket S.

Er variansen ukjent og n liten, bør en bruke $\bar{x} \pm t_{\alpha/2}(n-1)(s/\sqrt{n})$ til å beregne konfidensintervallet for μ. Her er $t_{\alpha/2}(n-1)$ definert slik at $P(T \geq t_{\alpha/2}(n-1)) = \alpha/2$, hvor T er t-fordelt med $n-1$ frihetsgrader. Intervallet er eksakt dersom $X_1, X_2, \ldots, X_n$ er normalfordelte, og vil ellers være en tilnærming. Tilnærmingen er generelt robust om ikke den underliggende fordelingen er sterkt asymmetrisk [22].

Dersom σ^2 er variansen til normalfordelte variable, er et tosidig konfidensintervall for σ^2 med konfidensnivå $100(1-\alpha)$ gitt ved

$$\left[\frac{(n-1)s^2}{b} \ , \ \frac{(n-1)s^2}{a} \right].$$

Her er $a = \chi^2_{1-\alpha/2}(n-1)$ og $b = \chi^2_{\alpha/2}(n-1)$. Kortere konfidensintervaller kan konstrueres [22], men dette vil ikke prioriteres her.

B Odds-ratio

Odds-ratio, OR, er oddsen for at en hendelse inntreffer i en gruppe dividert med oddsen for at den samme hendelsen inntreffer i en annen gruppe.

Matematisk defineres odds-ratio ved

$$\text{OR} = \frac{p/(1-p)}{q/(1-q)} = \frac{p(1-q)}{q(1-p)}.$$

Her er p sannsynligheten for at en hendelse inntreffer i den første gruppen, mens q er sannsynligheten for at hendelsen forekommer i den andre gruppen.

Dersom OR > 1, indikerer dette at hendelsen er mer sannsynlig i den første gruppen. Motsatt indikerer OR < 1 at hendelsen er mindre sannsynlig i denne første gruppen. Hvis OR $= 1$, er sannsynligheten for at hendelsen skal inntreffe lik i de to gruppene. Odds-ratioen er større enn eller lik null dersom denne er definert.

Odds-ratio brukes hyppig i case-control-studier. I forbindelse med metaanalyser måles effekten ofte i odds-ratio. Vi vet ikke sannsynligheten for at hendelsen, en suksess, skal inntreffe i de ulike gruppene. Denne sannsynligheten må derfor estimeres. Jeg lar $\sum_{i=1}^{n_1} Y_i$ betegne antall suksesser i n_1 uavhengige bernoulliforsøk i den første gruppen. Hvert forsøk har suksessannsynlighet p, slik at $P(Y_i = 1) = p$ og $P(Y_i = 0) = 1 - p$. Jeg definerer $\hat{p} = \sum_{i=1}^{n_1} Y_i/n_1$. Enkel regning gir at $\text{E}(\hat{p}) = p$ og at $\text{Var}(\hat{p}) = p(1-p)/n_1$. Estimatoren, $\hat{p}$, er forventningsrett for suksessannsynligheten. Sentralgrenseteoremet sikrer at $\hat{p}$ er asymptotisk normalfordelt.

På samme måte lar jeg $\sum_{i=1}^{n_2} X_i$ betegne antall suksesser i n_2 uavhengige bernoulliforsøk i den andre gruppen. Suksessannsynligheten er q. En forventningsrett estimator for q defineres ved $\hat{q} = \sum_{i=1}^{n_2} X_i/n_2$, hvor variansen er gitt ved $\text{Var}(\hat{q}) = q(1-q)/n_2$. Også $\hat{q}$ er asymptotisk normalfordelt.

Odds-ratioen estimeres ved uttrykket

$$\hat{\text{OR}} = \frac{\hat{p}(1-\hat{q})}{\hat{q}(1-\hat{p})}.$$

I denne oppgaven er jeg først og fremst interessert i log-odds-ratio. Log-odds-ratio defineres ved

$$\log(\text{OR}) = \log\left(\frac{p(1-q)}{q(1-p)}\right) = \log\left(\frac{p}{1-p}\right) - \log\left(\frac{q}{1-q}\right).$$

Jeg ønsker å utlede den asymptotiske fordelingen til log-odds-ratio-estimatoren,

$$\log(\hat{\text{OR}}) = \log\left(\frac{\hat{p}(1-\hat{q})}{\hat{q}(1-\hat{p})}\right) = \log\left(\frac{\hat{p}}{1-\hat{p}}\right) - \log\left(\frac{\hat{q}}{1-\hat{q}}\right).$$

Jeg definerer

$$\log(\hat{\text{OR}}) = \log\left(\frac{\hat{p}}{1-\hat{p}}\right) - \log\left(\frac{\hat{q}}{1-\hat{q}}\right) = f(\hat{p}) - f(\hat{q}).$$

Fordi $f'(p) = 1/\big(p(1-p)\big)$ og $f'(q) = 1/\big(q(1-q)\big)$, gir Deltametoden at

$$\sqrt{n_1}\left(\log\left(\frac{\hat{p}}{1-\hat{p}}\right) - \log\left(\frac{p}{1-p}\right)\right) \xrightarrow{L} N\left(0, \frac{1}{p(1-p)}\right)$$

og

$$\sqrt{n_2}\left(\log\left(\frac{\hat{q}}{1-\hat{q}}\right) - \log\left(\frac{q}{1-q}\right)\right) \xrightarrow{L} N\left(0, \frac{1}{q(1-q)}\right),$$

hvor $\xrightarrow{L}$ betegner konvergens i fordeling.

En lineær kombinasjon av to uavhengige, asymptotiske normalfordelte variable er også asymptotisk normalfordelt. Derfor har $\log\big((\hat{p}(1-\hat{q}))/(\hat{q}(1-\hat{p}))\big)$ en asymptotisk normalfordeling med forventning $\log\big(p(1-q)/(q(1-p))\big)$ og varians $1/\big(n_1 \cdot p(1-p)\big) + 1/\big(n_2 \cdot q(1-q)\big)$. Denne estimatoren har en additativ struktur. Dette medfører at $\log(\hat{\text{OR}})$ konvergerer raskere mot normalfordelingen enn $\hat{\text{OR}}$ som har en multiplikativ struktur [4]. For endelige sampelstørrelser er fordelingen til log-odds-ratio symmetrisk kun når forventningsverdien er lik null [6]. Figur 7, Seksjon 11.3, gir et inntrykk av den asymmetrien som oppstår for endelige sampelstørrelser.

Macaskill et al. [35] simulerer kontingenstabeller. Antall observerte suksesser og fiaskoer i behandlingsgruppen betegnes ved a_i og c_i henholdsvis. Tilsvarende er b_i og d_i antall observerte suksesser og fiaskoer i kontrollgruppen. Sannsynligheten for at en suksess inntreffer i behandlingsgruppen og kontrollgruppen estimeres derfor henholdsvis ved $a_i/(a_i + c_i)$ og $b_i/(b_i + d_i)$. Log-odds-ratio estimeres ved $\log(\hat{\text{OR}}) = \log\big((a_i d_i)/(b_i c_i)\big)$. Dette uttrykket er lik $\pm\infty$ hvis enten a_i, b_i, c_i eller d_i er lik 0. Uttrykket er ikke definert om både telleren og nevneren er lik null. Da disse mulighetene har positiv sannsynlighet, eksisterer ikke forventningen og variansen til $\log(\hat{\text{OR}})$ [4]. Estimatoren,

$$\log\left(\frac{(a_i + 0.5)(d_i + 0.5)}{(b_i + 0.5)(c_i + 0.5)}\right),$$

vil ha bedre egenskaper med hensyn til bias og MSE enn $\log(\hat{\text{OR}})$ [4]. Asymptotisk er effekten av å adderere 0.5 til hver celle neglisjerbar.

Fordi vi ikke har kjennskap til p eller q, estimeres variansen ved $\hat{\text{Var}}\big(\log(\hat{\text{OR}})\big) = 1/\big(n_1 \cdot \hat{p}(1-\hat{p})\big) + 1/\big(n_2 \cdot \hat{q}(1-\hat{q})\big)$. Med notasjonen innført hos Macaskill et al. [35] får vi at

$$\hat{\text{Var}}\big(\log(\hat{\text{OR}})\big) = \frac{a_i + c_i}{a_i c_i} + \frac{b_i + d_i}{b_i d_i} = \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}.$$

Variansestimatoren forbedres om vi også her adderer 0.5 til hver celle [4].

C Eksempel på simuleringskode brukt for å kontrollere Begg og Mazumdars simuleringsresultater

Her følger eksempel på simuleringskode brukt for å kontrollere Begg og Mazumdars simuleringsresultater [8] i situasjonen hvor vi har små metaanalyser, sterk seleksjonsstyrke og stor variansspredning. Vektfunksjonen avhenger av p-verdien for hypotesen om at den underliggende, sanne effekten er lik null. Denne koden bør leses i sammenheng med Kapittel 3, hvor formålet med simuleringene, samt utføringene, forklares inngående. Kun tekniske detaljer kommenteres i simuleringskoden.

```
# -- Deklarering av variable og tilordning av verdier --

rep <- 5000 # Antall ganger prosessen gjentas

antall_studier_gruppe1 <- 8 # Antall studier i gruppe 1
antall_studier_gruppe2 <- 9 # Antall studier i gruppe 2
antall_studier_gruppe3 <- 8 # Antall studier i gruppe 3
k <- antall_studier_gruppe1 + antall_studier_gruppe2 + antall_studier_gruppe3

# De ulike deltaverdiene (behandlingseffektene)
delta_vektor <- c(0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0)
l <- length(delta_vektor)

# Definerer vektfunksjonen
vektfunksjon <- function(a,b,p) exp(-b*p^a)

# Parametre som avgjor seleksjonsstyrken
a <- 1.5
b <- 4.0

epsilon <- 0.0001 # Unngaar ties vha vilkaarlig epsilon
nominelt_nivaa <- 0.05 # Tosidig nominelt nivaa

# Vektet gjennomsnittlig effektestimat
t_bar <- matrix(double(1),l,rep)

# Andelen genererte studier som inkluderes i metaanalysene
andel_inkludert <- matrix(double(1),l,rep)

# p-verdien for hypotesen om ingen publikasjonsbias
pverdi <- matrix(double(1),l,rep)

# Antall metaanalyser som inneholder signifikant publikasjonsbias
```

```
ant_forkast <- matrix(double(1),l,rep)

# -- Simuleringsdelen --

# For-lokke som gaar gjennom alle verdiene av delta
for (i in 1:l){

    # For-lokke som gjentar prosedyren rep antall ganger
    for (j in 1:rep){

        delta <- delta_vektor[i]

        # Varianser i hhv gruppe 1, 2 og 3
        v1 <- 10.0
        v2 <- 1.0
        v3 <- 0.1
        v <- v1

        # Tellere
        n <- 0
        s <- 0

        # Vektorer med variansene og de publiserte effektestimatene
        varians <- double(k)
        effektestimat <- double(k)

        # Simulerer metaanalyser

        # Simulerer antall_studier_gruppe1 publiserte studier med varians v1,
        # antall_studier_gruppe2 publiserte studier med varians v2 og
        # antall_studier_gruppe3 publiserte studier med varians v3

        while (s < k){

            # Antall studier som maa til for aa oppnaa k publiserte studier
            n <- n + 1

            if (s == antall_studier_gruppe1){
                v <- v2
            }
            if (s == antall_studier_gruppe1 + antall_studier_gruppe2){
                v <- v3
            }
```

```
    # Genererer effekt
    t <- rnorm(1,delta,sqrt(v))

    # Beregner p-verdi for hypotesen om at den underliggende effekten er lik null
    p <- pnorm(-t/(v^0.5))

    # Verdien til vektfunksjonen.
    # Denne settes lik 1 om en ikke onsker publikasjonsbias
    vf <- vektfunksjon(a,b,p)

    # Bruker binomisk modell for aa avgjore om studien
    # publiseres (x == 1) eller ikke (x == 0)
    x <- rbinom(1,1,vf)

    # Tar vare paa estimatene dersom studien publiseres
    if (x == 1){
        s <- s + 1
        effektestimat[s] <- t
        varians[s] <- v
        v <- v + epsilon
    }
}

# Standardiserer effektestimatene
t_bar[i,j] <- (sum(varians^(-1)*effektestimat))/sum(varians^(-1))
var_stjerne <- double(k)
t_stjerne <- double(k) # Standardisert effektestimat
for (m in 1:k){
    var_stjerne[m] <- varians[m]-(sum(varians^(-1)))^(-1)
    t_stjerne[m] <- (effektestimat[m]-t_bar[i,j])/(var_stjerne[m])^(0.5)
}

andel_inkludert[i,j] <- k/n

# Finner normalisert testobservator

# Ordner observasjonene
data <- data.frame(t_stjerne = t_stjerne, varians = varians)
ordnet_data <- data[order(data$t_stjerne),]

# Finner antall samsvarende par
c <- 0 # Samsvarende par
for (m in 1:k){
    for (w in m:k){
```

```
                if (ordnet_data$varians[m] < ordnet_data$varians[w]){
                    c <- c + 1
                }
            }
        }

        d <- 0.5*k*(k-1)-c # Ikke-samsvarende par

        # Normalisert testobservator
        z <- (c-d)/(((k*(k-1)*(2*k+5))/18))^0.5

        # Finner p-verdien for hypotesen om ingen publikasjonsbias
        if (z > 0){
            pverdi[i,j] <- (1 - pnorm(z))*2
        }
        if (z <= 0){
            pverdi[i,j] <- pnorm(z)*2
        }

        # Antall metaanalyser som inneholder signifikant publikasjonsbias
        if (pverdi[i,j] <= nominelt_nivaa){
            ant_forkast[i,j] <- ant_forkast[i,j] + 1
        }
    }
}

# Estimerer styrken
styrke <- rowMeans(ant_forkast)

# Beregner andelen simulerte studier som inkluderes i metaanalysen
andel_inkludert <- rowMeans(andel_inkludert)

# Biasen i estimatet for den underliggende effekten
bias <- double(l)
mean_t_bar <- rowMeans(t_bar)
for (i in 1:l){
    bias[i] <- mean_t_bar[i] - delta_vektor[i]
}

# -- Skriver ut de endelige resultatene --

endelige_resultater <- data.frame(styrke, andel_inkludert, bias)
endelige_resultater
```

D Eksempel på simuleringskode som viser hvordan nivået til rangkorrelasjonsmetodene kan tilpasses

Koden i dette vedlegget viser hvordan algoritmen i Seksjon 8.2 kan implementeres i praksis. Algoritmen forklarer en testmetode basert på rangkorrelasjoner hvor det virkelige signifikansnivået er tilnærmet lik det nominelle. I dette eksempelet brukes effektestimatene direkte. De standardiseres ikke.

```
# -- Deklarering av variable og tilordning av verdier --

# Vektoren med de aktuelle effektestimatene. Fyll inn
effektestimat <- c()

# Vektoren med de aktuelle variansene. Fyll inn
varians <- c()
k <- length(varians)

# Antall ganger simuleringene repeteres.
# Jo flere repetisjoner, desto mer noyaktige forkastningsintervaller
rep <- 10000

estimat_tau <- double(rep)

# -- Simuleringsdelen --

# Punkt 1 i algoritmen, Seksjon 8.2
for (i in 1:rep){
    # Genererer normalfordelte effektestimat ut fra de tilgjengelige variansene.
    # Uten tap av generalitet settes forventningen lik null.
    effektestimat_normal <- rnorm(k,0,sqrt(varians))

    # Korrelerer de nye effektestimatene og variansene vha Kendalls tau.
    # Kan eksempelvis nytte funksjonen cor.test() eller Kendall().
    estimat_tau[i] <- cor.test(effektestimat_normal, varians,
    method = "kendall")$estimate
}

# Punkt 2 i algoritmen
# Finner de onskede kvantilene til den empiriske fordelingsfunksjonen
forkastningsverdier <- quantile(estimat_tau, probs <- c(0.025,0.975))

# Punkt 3 i algoritmen
# Korrelerer effektestimatene fra den aktuelle metaanalysen og variansene
```

```
kendalls_tau <- cor.test(effektestimat, varians, method = "kendall")$estimate
kendalls_tau

# Punkt 4 i algoritmen
# Hvis kendalls_tau er mindre enn eller lik forkastningsverdier[1]
# eller storre enn eller lik forkastningsverdier[2],
# forkaster vi nullhypotesen om ingen publikasjonsbias.
# Hvis kendalls_tau ligger mellom forkastningsverdier[1]
# og forkastningsverdier[2], forkaster vi ikke denne nullhypotesen.

# Estimerer den tosidige mid-p-verdien
if (kendalls_tau < median(estimat_tau)){
    mid_p <- (length(which(estimat_tau < kendalls_tau)) +
    length(which(estimat_tau == kendalls_tau))/2)*2/length(estimat_tau)
}
if (kendalls_tau > median(estimat_tau)){
    mid_p <- (length(which(estimat_tau > kendalls_tau)) +
    length(which(estimat_tau == kendalls_tau))/2)*2/length(estimat_tau)
}
if (kendalls_tau == median(estimat_tau)){
    mid_p <- 1
}

mid_p
```

E Eksempel på simuleringskode brukt for å kontrollere simuleringsresultatene til Macaskill et al.

Her følger eksempel på simuleringskode brukt for å kontrollere simuleringsresultatene til Macaskill et al. [35] for Eggers uvektede regresjonsmetode, konfigurasjon A. Denne koden bør leses i sammenheng med Kapittel 11.1, hvor formålet med simuleringene, samt utføringene, forklares inngående. Kun tekniske detaljer kommenteres i simuleringskoden.

```
# -- Deklarering av variable og tilordning av verdier --

rep <- 10000 # Antall ganger prosessen repeteres

antall_gruppe1 <- 11 # Antall studier i gruppe 1
antall_gruppe2 <- 6 # Antall studier i gruppe 2
antall_gruppe3 <- 4 # Antall studier i gruppe 3
antall_studier <- antall_gruppe1 + antall_gruppe2 + antall_gruppe3

# Definerer vektfunksjonen
vektfunksjon <- function(alpha,beta,p) exp(-beta*p^alpha)

# Parametre som avgjor seleksjonsstyrken
alpha <- 1.5
beta <- 4

odds_ratio <- c(1, 2/3, 1/2, 1/4)
log_odds_ratio <- log(odds_ratio) # De underliggende behandlingseffektene
l <- length(odds_ratio)

theta_mh <- matrix(double(1),l,rep) # Mantel-Haenszels estimator
teller_mh <- matrix(double(1),l,rep) # Telleren i Mantel-Haenszels estimator
nevner_mh <- matrix(double(1),l,rep) # Nevneren i Mantel-Haenszels estimator

# Andelen studier som inkluderes i metaanalysen
andel_inkludert <- matrix(double(1),l,rep)

# Regresjonslinjens skjaeringspunkt
egger_estimat <- matrix(double(1),l,rep)

# Antall metaanalyser som inneholder signifikant publikasjonsbias
ant_forkast_egger <- matrix(double(1),l,rep) # Tosidig test
ant_forkast_egger_mindre <- matrix(double(1),l,rep) # Ensidig test
ant_forkast_egger_storre <- matrix(double(1),l,rep) # Andre hale
```

```
# -- Simuleringsdelen --

# For-lokke som gjennomgaar vektoren med de underliggende behandlingseffektene
for (i in 1:l){

    # For-lokke som gjentar prosedyren rep antall ganger
    for (j in 1:rep){

        # Vektorer med de publiserte effektestimatene og deres varianser
        effektestimat <- double(antall_studier)
        varians <- double(antall_studier)

        # Tellere
        k <- 0
        s <- 0

        # Simulerer metaanalyser hvor 11 publiserte studier
        # har 100 testgjenstander i baade kontroll- og behandlingsgruppen,
        # 6 publiserte studier har 200 testgjenstander i hver gruppe og
        # 4 publiserte studier har 300 i hver gruppe

        while (k < antall_studier){

            # Antall studier som maa genereres for aa oppnaa onsket antall publiserte studier
            s <- s + 1

            if (k < antall_gruppe1){
                # Antall testgjenstander i kontrollgruppen
                antall_kontroll <- 100
            }
            if (k >= antall_gruppe1 & k < antall_gruppe1 + antall_gruppe2){
                antall_kontroll <- 200
            }
            if (k >= antall_gruppe1 + antall_gruppe2 & k < antall_studier){
                antall_kontroll <- 300
            }

            # Antall testgjenstander i behandlingsgruppen
            antall_behandling <- antall_kontroll

            n <- antall_kontroll + antall_behandling

            # Suksessannsynligheten i kontrollgruppen
            suksessannsynlighet_kontroll <- runif(1,0.1,0.5)
```

```
# Finner suksessannsynligheten i behandlingsgruppen
b <- antall_kontroll * suksessannsynlighet_kontroll
d <- antall_kontroll - b
suksessannsynlighet_behandling <- b/d*odds_ratio[i]/(b/d*odds_ratio[i] + 1)

# Antall observerte suksesser i behandlingsgruppen
a <- rbinom(1, antall_behandling, suksessannsynlighet_behandling)

# Antall observerte suksesser i kontrollgruppen
b <- rbinom(1, antall_kontroll, suksessannsynlighet_kontroll)

# Antall observerte fiaskoer i behandlingsgruppen
c <- antall_behandling - a

# Antall observerte fiaskoer i kontrollgruppen
d <- antall_kontroll - b

# Adderer 0.5 til alle celler for aa redusere bias
a <- a + 0.5
b <- b + 0.5
c <- c + 0.5
d <- d + 0.5

# Beregner effektestimatet
t <- log((a * d)/(b * c))

# Beregner variansen
v <- 1/a + 1/b + 1/c + 1/d

# Beregner p-verdi for hypotesen om at den underliggende effekten er lik null
p <- pnorm(t/(v^0.5))

# Verdien til vektfunksjonen
# - settes lik 1 om en ikke onsker publikasjonsbias
vf <- vektfunksjon(alpha,beta,p)

# Bruker binomisk modell for aa avgjore om studien
# publiseres (x == 1) eller ikke (x == 0)
x <- rbinom(1,1,vf)

# Tar vare paa effektestimatet og variansen dersom studien publiseres
if (x == 1){
    k <- k + 1
```

```
        effektestimat[k] <- t
        varians[k] <- v

        # Oppdaterer teller og nevner i Mantel-Haenszels-estimatoren
        teller_mh[i,j] <- teller_mh[i,j] + ((a-0.5)*(d-0.5)/n)
        nevner_mh[i,j] <- nevner_mh[i,j] + ((b-0.5)*(c-0.5)/n)
    }
}

andel_inkludert[i,j] <- antall_studier/s

# Beregner Mantel-Haenszels estimator
theta_mh[i,j] <- teller_mh[i,j]/nevner_mh[i,j]

dep <- effektestimat/sqrt(varians) # Responsvariabelen i regresjonen
indep <- 1/sqrt(varians) # Forklaringsvariabelen i regresjonen
data <- data.frame(dep = dep, indep = indep)

# Eggers regresjon
egger <- lm(dep~indep, data <- data)

# Skjaeringspunktet
egger_estimat[i,j] <- summary(egger)$coef[1,1]

# p-verdi for hypotesen om ingen publikasjonsbias for tosidig test
egger_p_verdi <- summary(egger)$coef[1,4]

# Finner p-verdi for hypotesen om ingen publikasjonsbias
# for ensidig test og andre hale
t_verdi <- summary(egger)$coef[1,3]
egger_mindre_p_verdi <- pt(t_verdi, antall_studier - 2)
egger_storre_p_verdi <- 1 - pt(t_verdi, antall_studier - 2)

# Antall metaanalyser som inneholder signifikant publikasjonsbias.
# Tosidig nominelt nivaa paa 0.10
if (egger_p_verdi <= 0.10){
    ant_forkast_egger[i,j] <- ant_forkast_egger[i,j] + 1
}
if (egger_mindre_p_verdi <= 0.05){
    ant_forkast_egger_mindre[i,j] <- ant_forkast_egger_mindre[i,j] + 1
}
if (egger_storre_p_verdi <= 0.05){
    ant_forkast_egger_storre[i,j] <- ant_forkast_egger_storre[i,j] + 1
}
```

```
    }
}

# Styrkeestimat tosidig test
styrke_egger <- rowMeans(ant_forkast_egger)

# Styrkeestimat ensidig test
styrke_egger_mindre <- rowMeans(ant_forkast_egger_mindre)

# Styrkeestimat andre hale
styrke_egger_storre <- rowMeans(ant_forkast_egger_storre)

# Andelen genererte studier som inkluderes i metaanalysene
andel_inkludert <- rowMeans(andel_inkludert)

# Gjennomsnittlig skjaeringspunkt
gjennomsnitt_skjaeringspunkt <- rowMeans(egger_estimat)

# SE
standardfeil_skjaeringspunkt <- double(l)

# Biasen i estimatet for den underliggende effekten
bias <- double(l)

for (m in 1:l){
    standardfeil_skjaeringspunkt[m] <- sqrt(var(egger_estimat[m,])/rep)
    bias[m] <- mean(log(theta_mh[m,])) - log_odds_ratio[m]
}

# -- Skriver ut de endelige resultatene --

endelige_resultater <- data.frame(styrke_egger, styrke_egger_mindre, styrke_egger_storre,
gjennomsnitt_skjaeringspunkt, standardfeil_skjaeringspunkt, bias, andel_inkludert)
endelige_resultater
```

F Akseptert sammendrag i anledning ISCB 33

Improving the error rates of the Begg and Mazumdar test for publication bias in meta-analysis

Miriam Gjerdevik, Ivar Heuch
Department of Mathematics, University of Bergen, Bergen, Norway

The rank correlation test introduced by Begg and Mazumdar (1994) is widely used in meta-analysis to test for publication bias in clinical and epidemiological studies. It correlates the standardized treatment effect and the variance of the treatment effect.

However, it can be shown in simulations that the significance levels often deviate considerably from the nominal level. The assumptions for using the rank correlation test are not strictly satisfied. The pairs of observations fail to be independent, but the main cause of the poor significance level is a correlation between standardized effect sizes and sampling variances under the null hypothesis.

We propose alternative rank correlation tests to improve error rates. An unstandardized test directly correlates estimated effect sizes and sampling variances. This test reduces the Type II error rate, unfortunately at the expense of the Type I error rate. Simulations show that the standardized and unstandardized test statistics contain about the same amount of information. In tests for publication bias, it is essential to control the Type II error rate. If the significance level cannot be fixed, the unstandardized test is preferable.

Another test is based on the simulated distribution of the estimated measure of association, conditional on sampling variances. Its significance level equals the nominal level and the Type II error rate is reduced compared to the Begg and Mazumdar test. Although more computer intensive, this test attains the best significance levels.

Begg CB, Mazumdar M. (1994). Operating characteristics of a rank correlation test for publication bias. Biometrics 50, 1088-1102.

Referanser

[1] Mosby's medical dictionary. `http://medical-dictionary.thefreedictionary.com/meta-analysis`, 2009.

[2] Store norske leksikon. `http://snl.no/metaanalyse`, 2011.

[3] N. M. Adams, M. Crowder, D. J. Hand og D. Stephens. *Methods and models in statistics: in honour of Professor John Nelder, FRS*. London: Imperial College Press, 2004.

[4] A. Agresti. *Categorical data analysis*. New York: Wiley, 1990.

[5] C. B. Begg. Publication bias. I H. Cooper og L. V. Hedges, redaktører, *The handbook of research synthesis*, kapittel 25. New York: Russel Sage Foundation, 1994.

[6] C. B. Begg. Letter to the editor: A comparison of methods to detect publication bias in meta-analysis by P. Macaskill, S. D. Walter and L. Irwig, Statistics in Medicine, 2001; 20:641-654. *Statistics in Medicine*, 21:1803, 2002.

[7] C. B. Begg og J. A. Berlin. Publication bias: a problem in interpreting medical data. *Journal of the Royal Statistical Society, Series A*, 151:419–463, 1988.

[8] C. B. Begg og M. Mazumdar. Operating characteristics of a rank correlation test for publication bias. *Biometrics*, 50:1088–1101, 1994.

[9] D. J. Best og P. G. Gipps. Algorithm AS 71: the upper tail probabilities of Kendall's tau. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 23:98–100, 1974.

[10] B. Casella og R. L. Berger. *Statistical inference*. Pacific Grove, Calif.: Duxbury, andre utgave, 2002.

[11] R. Dersimonian og N. Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7:177–188, 1986.

[12] N. R. Draper og H. Smith. *Applied regression analysis*. New York: Wiley, tredje utgave, 1998.

[13] S. Duval og R. Tweedie. A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, 95:89–98, 2000.

[14] S. Duval og R. Tweedie. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56:455–463, 2000.

[15] M. Egger, G. D. Smith, M. Schneider og C. Minder. Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, 315:629–634, 1997.

[16] W. Feller. *An introduction to probability theory and its applications.* New York: Wiley, tredje utgave, 1968.

[17] C. Frost og S. G. Thompson. Correcting for regression dilution bias: comparison of methods for a single predictor variable. *Journal of the Royal Statistical Society, Series A*, 163:173–189, 2000.

[18] W. A. Fuller. *Measurement error models.* New York: Wiley, 1987.

[19] R. F. Galbraith. A note on graphical presentation of estimated odds ratio from several clinical trials. *Statistics in Medicine*, 7:889–894, 1988.

[20] I. Heuch. The distribution of the residual sum of squares in linear regression analysis with one predictor. Handout i faget STAT201, Universitetet i Bergen, Bergen, Norge, 2009.

[21] J. P. T. Higgins og S. G. Thompson. Quantifying heterogenety in a meta-analysis. *Statistics in Medicine*, 21:1539–1558, 2002.

[22] R. V. Hogg og E. A. Tanis. *Probability and statistical inference.* Upper Saddle River, N.J.: Pearson/Prentice Hall, syvende utgave, 2006.

[23] P. J. Huber. *Robust statistics.* New York: Wiley, 1981.

[24] L. Irwig, P. Macaskill, G. Berry og P. Glasziou. Bias in meta-analysis detected by a simple, graphical test. Graphical test is itself biased (Letter). *British Medical Journal*, 316:470, 1998.

[25] ISCB. *Program & abstract book, 33rd annual conference of the International Society for Clinical Biostatistics.* Bergen: ISCB, 2012.

[26] R. A. Johnson og G. K. Bhattacharyya. *Statistics: principles and methods.* Hoboken, N.J.: Wiley, femte utgave, 2006.

[27] M. Kendall og J. D. Gibbons. *Rank correlation methods.* London: Edward Arnold, femte utgave, 1990.

[28] W. H Kruskal. Ordinal measures of association. *Journal of the American Statistical Association*, 53:814–861, 1958.

[29] H. O. Lancaster. Significance tests in discrete distributions. *Journal of the American Statistical Association*, 56:223–234, 1961.

[30] L. Leng, T. Zhang, L. Kleinman og W. Zhu. Ordinary least square regression, orthogonal regression, geometric mean regression and their applications in aerosol science. *Journal of Physics: Conference series*, 78:84–89, 2007.

[31] R. J. Light og D. B. Pillemer. *Summing up: the science of reviewing research.* Cambridge, Mass.: Harvard University Press, 1984.

[32] J. Lillestøl. *Sannsynlighetsregning og statistikk: med anvendelser.* Oslo: Cappelen akademiske forlag, 1997.

[33] R. H. Lindeman, P. F. Merenda og R. Z. Gold. *Introduction to bivariate and multivariate analysis.* Glenview, Ill.: Scott, Foresman, 1980.

[34] G. G. Løvås. *Statistikk for universiteter og høyskoler.* Oslo: Universitetsforlaget, andre utgave, 2004.

[35] P. Macaskill, S. D. Walter og L. Irwig. A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine*, 20:641–654, 2001.

[36] A. I. McLeod. *Kendall: Kendall rank correlation and Mann-Kendall trend test.* R package version 2.2, 2011.

[37] K. Meen og I. Heuch. *Grensesetninger i sannsynlighetsregning: kompendium til emnet M251.* Bergen: Universitetet i Bergen, Matematisk institutt, 1984.

[38] R. B. Nelsen. *An introduction to copulas.* New York: Springer, 1998.

[39] NIST/SEMATECH. e-handbook of statistical methods. `http://www.itl.nist.gov/div898/handbook/`, 2012.

[40] G. E. Noether. *Elements of nonparametric statistics.* New York: Wiley, 1967.

[41] J. L. Peters, A. J. Sutton, D. R. Jones, K. R. Abrams og L. Rushton. Comparison of two methods to detect publication bias in meta-analysis. *The Journal of the American Medical Association*, 295:676–680, 2006.

[42] A. M. Pires og C. Amado. Interval estimators for a binomial proportion: comparison of twenty methods. *Statistical Journal*, 6:165–197, 2008.

[43] R Development Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2011.

[44] S. W. Raudenbush. Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: a synthesis of findings from 18 experiments. *Journal of Educational Psychology*, 76:85–97, 1984.

[45] Thomson Reuters. Web of knowledge. `m.webofknowledge.com/`, 2012.

[46] M. L. Rizzo. *Statistical computing with R.* Boca Raton, Fla.: Chapman & Hall/CRC, 2008.

[47] G. A. F. Seber. *Linear regression analysis.* New York: Wiley, 1977.

[48] D. J. Sheskin. *Handbook of parametric and nonparametric statistical procedures.* Boca Raton, Fla.: Chapman & Hall/CRC, tredje utgave, 2004.

[49] G. W. Snedecor og W. G. Cochran. *Statistical methods.* Ames, Iowa: Iowa State University Press, syvende utgave, 1980.

[50] N. Tang, Y. Wu, J. Ma, B. Wang og R. Yu. Coffee consumption and risk of lung cancer: a meta-analysis. *Lung Cancer*, 67:17–22, 2010.

[51] C. Thrane. *Regresjonsanalyse i praksis.* Kristiansand: Høyskoleforlaget, 2003.

[52] J. P. Vandenbroucke. Passive smoking and lung cancer: a publication bias? *British Medical Journal*, 296:391–392, 1988.

[53] W. N. Venables og B. D. Ripley. *Modern applied statistics with S.* New York: Springer, fjerde utgave, 2002.

[54] W. Viechtbauer. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36:1–48, 2010.

[55] S. E. Vollset. Confidence intervals for a binomial proportion. *Statistics in Medicine*, 12:809–824, 1993.

[56] [D. Walsh]. A derivation of Kendall's tau. `frank.mtsu.edu/~dwalsh/4370/437KTAU3.pdf`, 2006.

[57] F. Yates. Tests of significance for 2×2 contingency tables. *Journal of the Royal Statistical Society, Series A*, 147:426–463, 1984.