Stockholms universitet

Använder rekryterare föråldrade verktyg? En metaanalys av anställningsintervjuers validitet

Abir Gharbi

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2023:9
Matematisk statistik
Juni 2023

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2023:9**
http://www.math.su.se/matstat

Använder rekryterare föråldrade verktyg? En metaanalys av anställningsintervjuers validitet

Abir Gharbi*

Juni 2023

Sammanfattning

Syftet med denna metaanalys var att undersöka sambandet mellan arbetsintervjubedömningar och faktisk arbetsprestation. Analysen baserades på en samling av 160 studier med totalt 25 244 deltagare. Två inferensmetoder, frekventistisk och Bayesiansk, användes för att uppskatta sambandet, och resultaten jämfördes med en tidigare metaanalys utförd av McDaniel et al. (1994). De beräknade korrelationskoefficienterna erhållna från båda metoderna visade på en svagt positiv relation ($\hat{\rho}_F = 0.2325$, 95% KI: (0.2046, 0.2690); $\hat{\rho}_B = 0.2319$, 95% KrI: (0.2044, 0.2682)). Vid genomförandet av undergruppsanalyser framkom att situationsbaserade intervjuer hade en högre korrelation jämfört med andra intervjutyper. Dessutom påvisades att strukturerade intervjuer hade en högre korrelation än ostrukturerade.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: abgh5183@student.su.se. Handledare: Taras Bodnar, Kristoffer Lindensjö.

Abstract

This meta-analysis aimed to examine the relationship between job interview assessments and actual job performance. The analysis was based on a compilation of 160 studies involving a total of 25 244 participants. Two inference methods, frequentist and Bayesian, were used estimate the relationship, and the results were compared to a previous meta-analysis conducted by McDaniel et al. (1994). The estimated correlation coefficients obtained from both methods revealed a weak positive relationship ($\hat{\rho}_F = 0.2325$, 95% CI: (0.2046, 0.2690); $\hat{\rho}_B = 0.2319$, 95% CrI: (0.2044, 0.2682)). Subgroup analyses were conducted, indicating that situational interviews exhibited a higher correlation compared to other interview types. Additionally, structured interviews demonstrated a higher correlation than unstructured ones.

Förord

Detta är ett examensarbete på 15 högskolepoäng som leder till en kandidatexamen i matematisk statistik vid Stockholms universitet. Jag vill rikta ett stort tack till mina handledare Taras Bodnar och Kristoffer Lindensjö för deras värdefulla vägledning och stöd under arbetet med denna uppsats. Jag vill även tacka min vän Emma för hennes stöd och uppmuntran under hela arbetets gång.

Innehåll

1 Introduktion **5**
1.1 Syfte och mål 6

2 Teori **6**
2.1 Effektstorlek 6
2.1.1 Korrelation som effektstorlek 7
2.2 Statistiska modeller 8
2.2.1 Fixed effect-modellen 9
2.2.2 Random effects-modellen 9
2.2.2.1 Heterogeneitet 11
2.3 Modellval 12
2.4 Frekventistisk inferens och metaanalys 12
2.5 Bayesiansk inferens och metaanalys 14
2.5.1 A priorifördelning för μ och τ 17
2.5.2 Likelihood 18
2.5.3 A posteriorifördelning för μ och τ 20
2.5.4 DIRECT-metoden 21
2.6 Publiceringsbias 26

3 Data **28**
3.1 Beskrivning av variabler 29
3.2 Transformation av data 30

4 Metaanalys **31**
4.1 Frekventistisk 31
4.1.1 Undergruppsanalys 32
4.2 Bayesiansk 33
4.2.1 Undergruppsanalys 36
4.3 Publiceringsbias 38

5 Diskussion **40**

6 Slutsats **44**

1 Introduktion

I varje anställningsintervju hörs den fruktade öppningsfrasen, *"Berätta lite om dig själv"*, som verkar helt omöjlig att svara på varje gång - men kan denna enkla fråga verkligen förutsäga en persons lämplighet för ett jobb? Trots att anställningsintervjuer är vanlig praxis i anställningsprocessen har de bland annat kritiserats för sin subjektivitet och överdrivna förtroende för första intryck. Frågorna kan variera emellan intervjuare vilket leder till otydliga utvärderingskriterier, och kandidaterna kan svara på ett sätt som de tror att intervjuaren vill höra, istället för att uttrycka sina verkliga åsikter och färdigheter. Intervjumoment är även sårbara för fördomar som kan uppstå från omedvetna stereotyper. Trots detta fortsätter många organisationer att använda anställningsintervjuer för att bedöma potentiella anställda. Så frågan kvarstår: Är anställningsintervjuer verkligen ett effektivt verktyg för att bedöma en persons lämplighet för ett jobb?

Användningen av anställningsintervjuer har en lång historia som sträcker sig tillbaka till det tidiga 1900-talet. Tidigare fanns det varken en standardiserad norm eller effektiv teknologi för att söka arbete. Innan formella rekryteringsprocesser och online-ansökningar blev vanliga, var det vanligt att människor fick anställningar genom personliga kontakter, rekommendationer från vänner eller genom att direkt ta kontakt med potentiella arbetsgivare. Istället för att bli kallad till intervjuer var det vanligt att genomföra oplanerade besök hos arbetsgivare och informellt presentera sig själv och sin kompetens, med hopp om att säkra en anställning. Det var också vanligt att jobb gick i arv inom familjen och individer lärde sig ofta yrken eller färdigheter genom lärlingsutbildningar på jobbet. En ung person kunde lära sig att snickra genom att arbeta tillsammans med en mer erfaren snickare, en skogshuggare kunde anställa en lärling baserat på en referens från en pålitlig granne. När arbetsintervjuerna sedan infördes betonades ofta bedömningen av en individs fysiska och mentala lämplighet för en specifik befattning, snarare än att bedöma deras tidigare erfarenheter. Men i takt med att arbetsmarknaden utvecklades och blev mer komplex skedde samma sak med arbetsintervjuns roll.

Idag används intervjuer ofta för att bedöma en rad olika faktorer, exempelvis kompetens, kvalifikationer och personlighetsdrag. Även om det finns flera olika intervjuformat (t.ex. strukturerade, personlighetsbaserade, situationsbaserade), delar de alla det gemensamma målet att förse arbetsgivare med ett sätt att bedöma jobbkandidater. Trots dess omfattande användning har anställningsintervjuer länge kritiserats för orättvisa och partiskhet. Forskning har visat att intervjuare är mer benägna att anställa individer med likartade egenskaper som de själva (t.ex. kön, ras, ålder), eller som de uppfattar som mer socialt önskvärda (Rivera 2012, s. 1; Sears och Rowe 2003, s. 12). Men trots denna kritik fortsätter intervjuer att vara ett populärt verktyg i anställningsprocessen.

År 1994 genomförde McDaniel och hans kollegor en omfattande metaanalys om anställningsintervjuers validitet (McDaniel m. fl. 1994). Analysen inkluderade 160 oberoende studier och kombinerade resultaten för att sammanställa

intervjuers övergripande effektivitet som urvalsmetod. Deras resultat antydde att anställningsintervjuer var en svag, men ändå effektiv, prediktor för jobbprestation med en korrelation på 0.26. McDaniels arbete belyste även intervjuernas potentiella begränsningar, särskilt vad gäller deras mottaglighet för partiskhet och subjektiva bedömningar.

1.1 Syfte och mål

Med tanke på den omfattande användningen av anställningsintervjuer i anställningsprocessen är det av största vikt att undersöka deras giltighet som en prediktor för arbetsprestation. Syftet med detta arbete är att genomföra en metaanalys, med både frekventistiska och Bayesianska metoder, för att kombinera och analysera data från tidigare studier om validiteten av anställningsintervjuer. Dessa resultat kommer att jämföras med varandra samt med analysen utförd av McDaniel m. fl. 1994. Specifikt kommer denna uppsats att behandla följande frågor:

- Vad är korrelationen mellan anställningsintervju-bedömningar och faktisk jobbprestation? Skiljer den sig från den som är beräknad av Mcdaniel m.fl. (1994)?
- Skiljer sig resultaten mellan frekventistiska och Bayesianska metaanalyser? Vilka är styrkorna och begränsningarna hos de två tillvägagångssätten?
- Har intervjutyp och intervjustruktur någon större påverkan på korrelationen?

2 Teori

Metaanalys är en statistisk metod för att kombinera resultaten från tidigare gjorda studier, som oberoende av varandra undersökt samma eller liknande frågor. Syftet med en metaanalys är att ge en mer noggrann uppskattning av effekten av exempelvis en behandling eller åtgärd genom att sammanställa tidigare forskningsresultat (Hesser och Andersson 2015, s. 11). Metoden har blivit populär inom många områden, inklusive psykologi, medicin och ekonomi, på grund av dess förmåga att öka precisionen och generaliserbarheten av resultaten från enskilda studier (Borenstein m. fl. 2009, s. xxvi). Analysen slår inte bara samman data från mindre studier för att få en större stickprovsstorlek, utan den tar hänsyn till skillnader i studiestorlek, variation i hur studierna har genomförts och studiernas resultat.

2.1 Effektstorlek

Vanligtvis rapporteras resultaten från de enskilda studierna i form av en *effektstorlek* (eng. effect size). En effektstorlek är ett värde på styrkan hos ett samband mellan två variabler eller skillnaden mellan två grupper, och är i sin

tur den enhet som används i metaanalyser (Borenstein m. fl. 2009, s. 17). Det finns ett dussintal olika mått för effektstorlekar där de vanligaste är bland annat korrelation, standardiserad medelvärdesskillnad och oddskvoter (Tufanaru m. fl. 2015, s. 2). Ibland rapporteras dock inte resultaten på ett sätt som möjliggör för att extrahera den nödvändiga effektstorleksdatan från publicerade artiklar. Till exempel rapporteras ibland enbart resultaten av ett t-test, en envägsvariansanalys eller ett χ^2-test, men inte som mått såsom exempelvis korrelationer eller riskkvoter som är nödvändiga för en metaanalys. I sådana fall kan resultaten omvandlas till det format av effektstorlek som önskas. Det är även möjligt att omvandla mellan olika mått, vilket ofta görs då alla studier som valts till en metaanalys inte alltid rapporterar resultaten i samma format.

2.1.1 Korrelation som effektstorlek

När studier rapporterar resultat i form av korrelationer krävs det oftast ingen konvertering till något annat mått, utan korrelationskoefficienterna används som effektstorlek. Korrelation är ett intuitivt effektmått som är lätt att tolka och som samtidigt är standardiserat (Borenstein m. fl. 2009, s. 41). Med det menas att korrelation inte beror på de specifika skalor eller enheter som kanske använts för att mäta variablerna, vilket gör det enklare att jämföra styrkan hos samband mellan olika studier. Populationsparametern betecknas med ρ och skattningen av denna är stickprovskorrelationen r. Vidare har vi att variansen för r är approximativt (Borenstein m. fl. 2009, s. 41):

$$V_r = \frac{(1-\rho^2)^2}{n_i - 1}, \tag{1}$$

där n_i är antalet observationer i en studie i. I (1) ser vi att variansen starkt beror på korrelationen och detta gör det svårt att utföra metaanalyser direkt på korrelationskoefficienterna. Anledningen till att korrelationskoefficienterna inte används direkt är att stickprovsfördelningen av r kan vara skev (Kowalski 1972, s. 11). Detta påverkar variansen i fördelningen och i sin tur standardfelet för effektstorleken. Istället transformeras korrelationskoefficienterna till *Fishers* z vilket innebär att omvandla det ursprungliga korrelationsvärdet (r) till ett nytt värde (z) som följer en mer symmetrisk fördelning, som är mer önskvärt (Hesser och Andersson 2015, s. 69). Fishers z-transformation används för att stabilisera variansen V_r och för att beräkna ett konfidensintervall (KI) för korrelationskoefficienten, speciellt i analyser av mindre stickprovsstorlekar. Denna transformation ges av (Borenstein m. fl. 2009, s. 42)

$$z = 0.5 \cdot ln\left(\frac{1+r}{1-r}\right). \tag{2}$$

Variansen av z ges (till en god approximation) av

$$V_z = \frac{1}{n_i - 3} \tag{3}$$

och standardfelet är därmed

$$SE_z = \sqrt{V_z} \tag{4}$$

(Borenstein m. fl. 2009, s. 42). Här noterar vi att variansen (och standardfelet) inte längre beror på den underliggande korrelationen, utan endast på stickprovsstorleken n_i hos en studie. Från formel (3) ser vi att när antalet deltagare/observationer ökar i stickprovet kommer variansen att minska, vilket leder till att standardfelet i sin tur blir mindre. Det beräknade standardfelet kan användas för att tilldela en vikt till en enskild studie samt för att konstruera konfidensintervall för effektstorleken i den studien. Konfidensintervall blir ett sätt att visa inom vilket intervall vi med en viss säkerhet kan säga att den sanna populationsparametern ρ befinner sig i.

Under metaanalysen används alltså inte variansen för r men istället används Fishers z-värdet och dess varians. Varje studie har ett korrelationsvärde r som sedan omvandlas till z i analysen för att ge en sammanfattande effekt, konfidensintervall och så vidare i Fishers z-mått (Borenstein m. fl. 2010, s. 42). Sedan konverteras resultaten av metaanalysen tillbaka till korrelationsenheter med hjälp av

$$r = \frac{e^{2z} - 1}{e^{2z} + 1}. \tag{5}$$

2.2 Statistiska modeller

De flesta metaanalyser baseras på en av två statistiska modeller; fixed effect modellen (FEM) eller random effects-modellen (REM). Under fixed effect-modellen antas att det finns *en* gemensam effekt som ligger till grund för alla studier i metaanalysen, och att alla skillnader i observerade effekter beror på stickprovsfel (Borenstein m. fl. 2010, s. 61). Med andra ord antas det att alla studier dragits från samma population och att det är en sann effekt som gäller för denna population.

För random effects-modellen antas det att den sanna effekten kan variera från studie till studie, alltså att studierna har dragits från olika populationer. Exempelvis kan effekstorleken variera mellan studier där deltagarna är av olika åldrar, olika sjuka eller där andra metoder/verktyg har använts. En annan faktor som kan påverka är om en intensivare variant av behandling/åtgärd använts i vissa studier, som exempelvis en starkare version av en medicin eller ett striktare kostschema. (Borenstein m. fl. 2009, s. 61) Eftersom studierna kan skilja sig åt i deltagare och implementering av intervention, bland andra skäl, går det inte längre att anta att det finns en gemensam effekt då effekten troligtvis är olika i de olika studiepopulationerna. Om det vore möjligt att genomföra ett oändligt antal studier skulle de sanna effektstorlekarna för dessa studier fördelas kring något medelvärde (Dettori, Norvell och Chapman 2022, s. 1). Effektstorlekarna i de faktiskt genomförda studierna antas representera ett slumpmässigt urval av de oändligt många studierna, därav namnet *random* effects-modellen. Man kan se det som att REM skattar medelvärdet över alla ingående populationer.

2.2.1 Fixed effect-modellen

Under FEM antas att alla effektstorlekar härrör från en enda homogen population och att alla studierna i analysen har en gemensam (sann) effektstorlek. Denna sanna och okända effekt är den totala effektstorleken vi vill beräkna i vår metaanalys och vi betecknar den med θ. Den observerade effektstorleken för en studie $i = 1, 2, ..., k$ betecknas med Y_i.

Enligt modellen är den enda anledningen till att en studies observerade effektstorlek Y_i avviker från θ på grund av dess stickprovsfel ε_i. Under FEM menas att alla studier är skattningar av samma egentliga effektstorlek. Men eftersom varje studie bara kan dra något större eller mindre urval ur den oändligt stora studiepopulationen, belastas resultaten av stickprovsfel. Detta stickprovsfel gör att den observerade effekten avviker från den övergripande, sanna effekten. Mer generellt ges den observerade effekten Y_i för varje studie av populationsmedelvärdet plus stickprovsfelet i den studien. Vi kan beskriva detta samband enligt (Borenstein m. fl. 2009, s. 64–65):

$$Y_i = \theta + \varepsilon_i, \quad \varepsilon_i \overset{ind}{\sim} \mathcal{N}(0, \sigma_i^2) \tag{6}$$

Stickprovsfelen ε_i antas vara normalfördelade med medelvärde noll och varians $\sigma_i^2 > 0$, där "ind" indikerar att de slumpmässiga felen är oberoende av varandra. Även om vi inte känner till den sanna effektstorleken θ för våra studier kan vi utnyttja detta samband för att beräkna den bästa möjliga skattningen av θ. Den sanna effektstorleken skattas med den viktade genomsnittliga effektstorleken för alla studier. Likaså är ε_i okänd och skattas med hjälp av standardfelet för varje studie. Kortfattat har vi att de viktigaste antagandena för FEM är:

1. Homogenitet av effektstorlekar: Den sanna effektstorleken är densamma för alla studier.

2. Oberoende: Varje studie är oberoende av de andra.

3. Normalfördelade stickprovsfel: Den slumpmässiga felkomponenten ε_i i varje studie är normalfördelad med medelvärde noll och varians σ_i^2.

4. Fixt effekter: Den verkliga effektstorleken är fixt i alla studier.

Om dessa antaganden inte uppfylls kan det hända att FEM inte ger korrekta uppskattningar av den verkliga effektstorleken.

2.2.2 Random effects-modellen

Den tidigare diskuterade modellen med fixt effekt utgår från antagandet att den sanna effekten är densamma över alla studier. Detta antagande kan dock vara orimligt. När man väljer en grupp av studier att inkludera i en metaanalys antas det att studierna har tillräckligt stora likheter för att det ska vara rimligt att sammanställa informationen. Det finns däremot generellt sett ingen anledning att anta att de är "identiska" i den meningen att den sanna effektstorleken är

exakt densamma i alla studier. Studier skiljer sig oftast åt, även om det bara är på knappt märkbara sätt. Responsvariabeln som är av intresse kan ha mätts på olika sätt eller på olika platser. Kanske var inte typen av behandling exakt densamma eller att behandlingens intensitet och längd skilde sig. Det är inte ovanligt att målpopulationen för studierna inte var identisk eller att det fanns skillnader i de kontrollgrupper som användes.

Det är även troligt att studierna som ingår metaanalysen inte bara varierar på en utan flera aspekter samtidigt. Om så är fallet kan vi förvänta oss en betydande *heterogenitet* (varians) mellan studierna. Ett sätt att hantera variationen mellan studier är att genomföra en analys med REM. Under REM antas vanligtvis att de sanna effektstorlekarna är normalfördelade (Borenstein m. fl. 2009, s. 70). Modellen ger då en skattning av medelvärdet över alla olika effektstorlekar från studierna, och inkluderar även en komponent för att fånga upp den "sanna" variationen i effektstorlekar i populationen, utöver slumpfel.

Med andra ord antar REM att en observerad effektstorlek Y_i är en skattning av studiens sanna effektstorlek θ_i, som påverkas av stickprovsfel ε_i:

$$Y_i = \theta_i + \varepsilon_i. \tag{7}$$

Faktum att θ_i används istället för θ visar på en viktig skillnad mellan modellerna. REM skiljer sig från FEM genom att anta att θ_i endast är den sanna effektstorleken för studie i och antagandet menar att det finns en annan källa till fel, betecknad med ξ_i. Denna andra felkällan introduceras av det faktum att även den sanna effektstorleken θ_i, för studie i, endast utgör en *del* av en övergripande fördelning av sanna effektstorlekar, med medelvärdet μ (Borenstein m. fl. 2009, s. 71):

$$\theta_i = \mu + \xi_i. \tag{8}$$

Därför beskriver REM en hierarki av två processer som inträffar i analysen. Observerade effektstorlekar Y_i av en studie avviker från sitt sanna värde θ_i på grund av stickprovsfel ε_i. Samtidigt är de sanna effektstorlekarna θ_i endast ett urval från ett universum av sanna effekter, vars medelvärde (μ) vi vill uppskatta som den genomsnittliga effekten av vår metaanalys. Genom att ersätta θ_i i (7) med dess definition i (8) kan vi uttrycka random effects-modellen som (Borenstein m. fl. 2009, s. 71)

$$Y_i = \mu + \xi_i + \varepsilon_i. \tag{9}$$

Denna ekvation gör det tydligt att Y_i avviker från den samlade effektstorleken μ på grund av två feltermer, ξ_i och ε_i.

Sammanfattningsvis antar REM att det finns heterogenitet bland de sanna effektstorlekarna för studierna och målet är att uppskatta den samlade effektstorleken μ.

Kortfattat har vi att de viktigaste antagandena för REM är:

1. Heterogeneitet av effektstorlekar: Den sanna effektstorleken är olika för alla studier men följer samma fördelning.

2. Oberoende: Varje studie är oberoende av de andra. Det gäller även att alla stokastiska variabler ξ_i och ε_i är oberoende.

3. Normalfördelade fel: Det gäller att

   (a) $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$

   (b) $\xi_i \sim \mathcal{N}(0, \tau^2)$, där τ^2 är variansen mellan studier.

4. Slumpmässiga effekter: Den sanna effektstorleken är slumpmässig i alla studier.

2.2.2.1 Heterogeneitet

Utmaningen förknippad med REM är att vi måste ta hänsyn till feltermen ξ_i. För att göra detta måste vi uppskatta variansen i fördelningen av de sanna effektstorlekarna θ_i. Denna varians är känd som τ^2 och är variansen, även kallad heterogeneiteten, *mellan* studierna. Med andra ord, om vi på något sätt visste den sanna effektstorleken för varje studie och beräknade variansen för dessa effektstorlekar hade denna varians varit τ^2 (Borenstein m. fl. 2009, s. 72). Det finns inga tydliga tumregler för vad ett högt eller lågt τ^2-värde är, utan det beror i hög grad på den specifika frågan som metaanalysen svarar på eller hur analysen är utformad (t.ex. vilka studier som ingår). Detta gör att tolkning av τ^2 kan vara svår och att det därav kan vara nyttigt att använda τ^2 tillsammans med andra mått på heterogeneiteten. Ytterligare en anledning till varför τ^2 inte bör vara den enda statistikan som används för att tolka heterogeneitet är för att τ^2 antar att effektstorlekarna är normalfördelade, vilket inte alltid är fallet. Dessutom ger måttet ingen information om heterogeneitetens riktning eller natur.

Ett annat sätt att mäta heterogeneiteten på är med *Cochrans* Q-test som är ett icke-parametriskt test vilket τ^2 bygger på (Borenstein m. fl. 2009, s. 109). Cochrans Q är χ^2-fördelad med $k-1$ frihetsgrader där k är antalet studier. Under antagandet om nollhypotesen, vilket innebär ingen heterogenitet, är den förväntade mellan-studier-variansen i effektstorlekarna noll. Om det beräknade Q-värdet överstiger det kritiska värdet av en χ^2-fördelning vid en given signifikansnivå (t.ex. 0.05) kan slutsatsen dras att det finns signifikant heterogeneitet mellan studierna.

Den vanligt förekommande I^2-statistikan ger en alternativ metod för att kvantifiera heterogeneiteten och härleds direkt från Cochrans Q (Higgins, Thompson m. fl. 2003, s. 2). Statistikan mäter den procentuella variationen i effektstorlekarna som inte beror på stickprovsfel och är därför ett intuitivt mått då den ger värden mellan 0% och 100%. Higgins och Thompson har föreslagit en inofficiell klassificering av I^2-värden med syfte att hjälpa till att tolka dess magnitud. Således skulle procentandelar på cirka 25% ($I^2 = 25$), 50% ($I^2 = 50$) och 75% ($I^2 = 75$) innebära låg, måttlig respektive hög heterogeneitet (Higgins, Thompson m. fl. 2003, s. 3).

Sammanfattningsvis ger Cochrans Q-test en prövning av heterogeneitet, τ^2 uppskattar mängden av den och I^2 beskriver andelen av total variation i effektstorelkarna som beror på heterogeneitet.

2.3 Modellval

Valet mellan de två modellerna beror bland annat på forskningsfrågan och de antaganden som är mest lämpliga för datan. Generellt förderas ofta REM över FEM då den är mer "realistisk" och tar hänsyn till heterogeneitet mellan studierna (Dettori, Norvell och Chapman 2022, s. 2). Dessutom är FEM den lämpliga modellen när antalet studier är litet och REM när antalet studier är tillräckligt stort, det vill säga tillräckligt många studier för att kunna stödja generella slutsatser utöver de inkluderade studierna (Tufanaru m. fl. 2015, s. 7). Vi kommer i detta arbete fokusera på och använda oss utav REM och följande avsnitt redogör för hur modellens parametrar skattas på två olika tillvägagångssätt – frekventistiskt och Bayesianskt.

2.4 Frekventistisk inferens och metaanalys

Det frekventistiska tillvägagångsättet är en typ av statistisk inferensteori som testar om en händelse (hypotes) inträffar. Teorin behandlar sannolikheter som ekvivalenta med "frekvenser" - antalet gånger någonting händer. Idén bakom frekventistisk inferens är att använda samlad data för att skatta parametrarna av intresse och att utvärdera noggrannheten hos dessa skattningar med hjälp av sannolikhetsfördelningar och statistiska tester. I frekventistisk metaanalys är målet att skatta populationseffekten genom att använda de observerade effekterna. För att få den mest exakta skattningen av μ beräknar vi ett viktat medelvärde, där vikten som tilldelas varje studie är inversen av studiens varians. Under FEM är variansen (och därav vikten) för en studie i endast beroende av stickprovsstorleken, minns formel (3):

$$V_{Y_i} = \frac{1}{n_i - 3} \quad \text{resp.} \quad W_i = n_i - 3. \tag{10}$$

där V_{Y_i} är variansen inom studie i och W_i är vikten tilldelad till studien.

För REM är variansen för en studie summan av V_{Y_i} och τ^2. Vi låter $\hat{\tau}^2$ beteckna skattningen av τ^2 och får att vikten som tilldelas till varje studie är (Borenstein m. fl. 2009, s. 73):

$$W_i^* = \frac{1}{V_{Y_i}^*}, \tag{11}$$

där V_{Y_i} är variansen inom en studie i plus variansen mellan studierna, $\hat{\tau}^2$. Det vill säga

$$V_{Y_i}^* = V_{Y_i} + \hat{\tau}^2.$$

Här betecknar asterisken (*) att beräkningen gäller random effects-modellen. Det finns flera metoder för att skatta τ^2 där en av de vanligare är DerSimonian

and Laird-metoden (Borenstein m. fl. 2009, s. 72). Den beräknar $\hat{\tau}^2$ enligt

$$\hat{\tau}^2 = \frac{Q - df}{C}, \tag{12}$$

där Q är Cochrans Q och beräknas enligt

$$Q = \sum_{i=1}^{k} W_i Y_i^2 - \frac{\left(\sum_{i=1}^{k} W_i Y_i\right)^2}{\sum_{i=1}^{k} W_i}, \tag{13}$$

$$df = k - 1,$$

och

$$C = \sum_{i=1}^{k} W_i - \frac{\sum_{i=1}^{k} W_i^2}{\sum_{i=1}^{k} W_i}. \tag{14}$$

Notera att vikten W_i är utan asterisk (*). Under REM ges statistikan I^2 av (Higgins och Thompson 2002, s. 6):

$$I^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \tilde{v}}, \tag{15}$$

där

$$\tilde{v} = \frac{(k-1)\sum_{i=1}^{k} W_i}{(\sum_{i=1}^{k} W_i)^2 - \sum_{i=1}^{k} W_i^2}. \tag{16}$$

Skattningen $\hat{\mu}$ av μ ges då av det viktade medelvärdet

$$\hat{\mu} = \frac{\sum_{i=1}^{k} W_i^* Y_i}{\sum_{i=1}^{k} W_i^*}, \tag{17}$$

variansen estimeras enligt

$$V_{\hat{\mu}^*} = \frac{1}{\sum_{i=1}^{k} W_i^*} \tag{18}$$

och standardfelet

$$SE_{\hat{\mu}^*} = \sqrt{V_{\hat{\mu}^*}}. \tag{19}$$

Random effects-modellen syftar alltså till att uppskatta medelvärdet av en fördelning av effektstorlekar, snarare än en sann effekt. Eftersom varje studie ger information om olika effektstorlekar är det avgörande att inkludera *alla* effektstorlekar i den sammanfattande uppskattningen. Med det menas att mindre studier inte kan bortses från genom att tilldela dem en mycket liten vikt, som man hade gjort under FEM. Även om studien och dess resultat inte är tillräckligt noggranna ger uppskattningen från en liten studie information om en effektstorlek som ingen annan studie har uppskattat. På samma sätt kan vi inte tilldela en stor studie för mycket vikt. Målet är att uppskatta medeleffekten över en rad olika studier och vi måste säkerställa att den sammanfattande uppskattningen inte påverkas alltför mycket av någon studie. Nollhypotesen som testas i detta arbete är om den sammanfattande medelefffekten är noll, alltså

$$H_0 : \mu = 0.$$

2.5 Bayesiansk inferens och metaanalys

Den frekventistiska metoden uppskattar okända parametrar, bestämmer konfidensintervall och utför statistiska tester för att avgöra i vilken utsträckning resultaten är förenliga med en nollhypotes (genom att undersöka p-värden). Ett alternativ till det tillvägagångssättet är den Bayesianska metoden/läran. Denna lära bygger på en annan sannolikhetsfilosofi där sannolikhet tolkas som en osäkerhet eller tro, istället för en frekvens. Bayesiansk statistik grundar sig i Bayes sats som låter oss uppskatta sannolikheten för en händelse A, givet att vi redan vet att en annan händelse B har inträffat. Detta resulterar i en betingad sannolikhet som betecknas $P(A|B)$. Satsen beskrivs enligt följande

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}, \tag{20}$$

där $0 < P(A) < 1$ och $P(B) > 0$ (Held och Sabanés Bové 2014, s. 168). Satsen beskrivs oftast som den betingade sannolikheten av två händelser A och B, men vi kan även se A och B som sannolikhetsfördelningar för två variabler. I Bayesiansk inferens och metaanalys är parametern θ som vi vill studera stokastisk med en a priori-fördelning $f(\theta)$. Efter att ha observerat data Y (som också följer en viss fördelning) kan Bayes sats implementeras för att beräkna a posteriori-fördelningen $f(\theta|Y)$. För två stokastiska variabler Y och θ kan den betingade sannolikheten skrivas som (Held och Sabanés Bové 2014, s. 320)

$$f(Y|\theta) = \frac{f(Y,\theta)}{f(\theta)}. \tag{21}$$

En omskrivning av denna ekvation ger den simultana sannolikheten (eng. *joint probability*)

$$f(Y,\theta) = f(Y|\theta)f(\theta)$$

och det gäller även att

$$f(Y,\theta) = f(\theta|Y)f(Y).$$

Vi kan nu med detta härleda täthetsfunktionen för a posteriorifördelningen (Held och Sabanés Bové 2014, s. 170):

$$f(\theta|Y) = \frac{f(Y|\theta)f(\theta)}{f(Y)} = \frac{f(Y|\theta)f(\theta)}{\int f(Y|\theta)f(\theta)d\theta}. \tag{22}$$

Termen $f(Y|\theta)$ i ekvationerna ovan beskriver likelihood-funktionen och den ger oss sannolikheten att observera Y givet θ. Vi har att nämnaren i (22) kan skrivas om som

$$\int f(Y|\theta)f(\theta)d\theta = \int f(Y,\theta) = f(Y),$$

där $f(Y)$ är den marginella fördelningen för datan och som samtidigt inte beror på θ. Detta innebär att a posteriori-fördelningen kan skrivas om som

$$f(\theta|Y) \propto f(Y|\theta)f(\theta). \tag{23}$$

Notera att högerledet i (23) inte längre är en väldefinierad täthetsfunktion (då den inte integreras till ett) men den är proportionell mot a posteriorifördelningen, vilket innebär att de två ekvationerna (22) och (23) skiljer sig åt med en proportionalitetskonstant $\frac{1}{f(Y)}$. Därav kommer vissa skattningar att vara desamma, såsom medelvärdet, medianen och kredibilitetsintervall.

A posterori-medelvärdet $E(\theta|Y)$ är väntevärdet för a posteriorifördelningen och definieras enligt (Held och Sabanés Bové 2014, s. 171)

$$E(\theta|Y) = \int \theta f(\theta|x) d\theta.$$

A posterori-medianen $Med(\theta|Y)$ är medianen för a posteriorifördelningen, d.v.s. ett godtyckligt tal a så att

$$\int_{-\infty}^{a} (\theta|Y) d\theta = 0.5 \quad \text{och} \int_{a}^{\infty} (\theta|Y) d\theta = 0.5.$$

Vi kan säga att Bayesiansk statistik "fäster" sannolikhetsfördelningar till parametrarna av intresse, såsom μ och τ. Exempelvis kan vi representera vår osäkerhet om μ för en samling studier genom att välja t.ex. en normalfördelning centrerad på vår *bästa gissning*, med svansar som beskriver hur säkra vi är på vår gissning. Analysen börjar alltså med att "gissa" en a priori-sannolikhetsfördelning till varje okänd kvantitet, vilket beskriver en a priori osäkerhet eller tro om kvantiteten innan man ser datan.

Bayesiansk analys kombinerar a priorifördelningen med datan för att producera en a posteriorifördelning för den okända kvantiteten. Ibland kan en a priorifördelning representera den osäkerhet man har istället för en gissning, med exempelvis en likformig fördelning. När a priorifördelningen representerar osäkerhet är a posteriorifördelningen en sammanfattning av vad datan säger oss om parametern i fråga. Man kan se det Bayesianska sättet som en generalisering av den frekventistiska, där man kan inkludera förhandsinformation om så önskas.

Vidare har vi att Bayesianska analyser producerar kredibilitetsintervall (KrI, eng. *credible interval*) istället för konfidensintervall, vilket sammanfattar a posteriori-fördelningen och uttrycker sannolikheten att den sanna kvantiteten är inom intervallet (Held och Sabanés Bové 2014, s. 172). Ett kredibilitetsintervall definieras av två värden, t_l och t_u, som motsvarar nedre och övre gränsen. För ett fixt $\gamma \in (0, 1)$ definieras ett $\gamma \cdot 100\%$ kredibilitetsintervall enligt

$$\int_{t_l}^{t_u} f(\theta|Y) d\theta = \gamma, \tag{24}$$

där γ kallas kredibilitetsnivån (eng. *credible level*).

Med hjälp av detta kan vi också beräkna sannolikheten att en uppskattning är större eller mindre än noll. Till skillnad från att använda p-värden kräver detta uttalande ingen nollhypotes för dess tolkning!

En av utmaningarna med Bayesiansk inferens är att felaktig a priori-information kan leda till slutsatser i felaktig riktning (Carlin och Louis 2000, section

4.4.1). Samtidigt finns det inget objektivt "korrekt" sätt att välja en a priorifördelning; det kräver snarare en viss skicklighet att översätta subjektiva a priori-idéer till matematiskt formulerade fördelningar. Vanligtvis jämförs flera olika a priorifördelningar med varanda och om väldigt olika fördelningar leder till samma a posteriorifördelning kan analysen betraktas som *robust*. Men om a priorifördelningen istället är *inflytelserik* kan detta indikera att analysen baseras på otillräcklig data. I sådana fall bör både Bayesianska och frekventistiska metaanalyser tolkas med försiktighet.

För att genomföra en Bayesiansk metaanalys tillämpar vi en Bayesiansk hierarkisk modell. Denna modell är besläktad med random effects-modellen (REM) men uttrycks på ett annorlunda sätt. Med begreppet "hierarkisk" avses att modellen innefattar flera *nivåer* av variation eller hierarki i data, där varje nivå har sina egna uppsättningar av parametrar. Detta gör det möjligt att hantera mer komplexa strukturer där osäkerhet kan inkorporeras på varje nivå. På den första nivån har vi deltagarna i studien och data på denna nivån är oftast i formen av beräknade effektstorlekar Y_i för varje enskild studie i. På den andra nivån antar vi att deltagarna är nästlade inom studierna, till exempel grupper indelade efter åldrar eller utbildningsnivå, och att de sanna effektstorlekarna θ_i för olika studier följer en egen fördelning. Denna fördelning av sanna effektstorlekar har ett medelvärde μ och en varians τ^2, som återigen representerar heterogeneiteten mellan studier.

Hierarkin kan beskrivas så här.

1. Första steget: Här antar vi att den observerade effektstorleken Y_i är en skattning av den sanna effekten θ_i för studie i. Den observerade effekten Y_i avviker från θ_i p.g.a. stickprovsfelet ε_i, detta då vi antar att Y_i är tagen ur populationen som är underliggande studien. Denna population kan ses som en fördelning med medelvärde θ_i (den sanna effekten hos studien) och varians σ_i^2.

2. Andra steget: Här antar vi att de sanna effektstorlekarna θ_i själva är stickprov från en övergripande fördelning. Medelvärdet μ av denna fördelning är den sammanfattade effektstorleken vi vill uppskatta. De sanna effekterna θ_i avviker från μ då den övergripande föredelningen också har en varians, nämligen mellan-studie heterogeneiteten τ^2.

Tillsammans ger stegen dessa uttryck:

$$Y_i \sim \mathcal{N}(\theta_i, \sigma_i^2), \quad \theta_i \sim \mathcal{N}(\mu, \tau^2) \tag{25}$$

Modellen kan förenklas och vi får

$$Y_i \sim \mathcal{N}(\mu, \sigma_i^2 + \tau^2). \tag{26}$$

Dessa uttryck/ekvationer är väldigt lika de som diskuterades under REM-avsnittet och det finns egentligen inget "Bayesianskt" med dem. Men detta ändras när vi lägger till följande ekvationer (Williams, Rast och Bürkner 2018,

s. 4):

$$\begin{aligned} (\mu, \tau^2) &\sim p(.) \\ \tau^2 &> 0. \end{aligned} \tag{27}$$

Här är den första raden i (27) av särskilt intresse. Den definierar a priorifördelningar för parametrarna μ och τ^2 och tillåter oss att specificera a priori vad vi tror om parametrarna och hur säkra vi är på dessa antaganden. Den andra raden inför en begränsning som kräver att heterogeniteten mellan studierna är större än noll. Det bör dock noteras att dessa rader inte specificerar den **typ** av a priorifördelning som ska användas, utan endast antyder att **någon form** av a priorifördelning antas.

2.5.1 A priorifördelning för μ och τ

Vi har etablerat den hierarkiska modellen som kan användas för utföra en Bayesiansk metaanalys. Men för att modellen ska fungera måste vi specificera a priorifördelningarna för μ och τ (Röver 2020, s. 7). Det är avgörande att välja dessa fördelningar klokt, särskilt när antalet studier är litet, eftersom a priorifördelningarna kan ha en betydande inverkan på resultaten. När det gäller skattningen av μ rekommenderas generellt att använda *svagt informativa* a priorifördelningar, men även *icke-informativa* a priorifördelningar förekommer också (Williams, Rast och Bürkner 2018, s. 13). Icke-informativa a priorifördelningar, som utgör den enklaste formen av fördelningar, bygger på den likformiga fördelningen och antar att alla värden är lika trovärdiga. Å andra sidan är svagt informativa a priorifördelningar mer avancerade och bygger på fördelningar som återspeglar osäkerheten kring a priori-antaganden. Dessa fördelningar är oftast att föredra, även om de inte heller ger några specifika uttalanden om värdet på μ. Ett naturligt val av en svagt informativ a priorifördelning för μ är normalfördelningen, med medelvärde μ_p och varians σ_p^2 (Röver 2020, s. 8). Detta arbete kommer att fokusera på dessa två nämnda a priorifördelningar för μ, det vill säga likformig och normal fördelning.

Nästa a priorifördelning som ska bestämmas är den för τ. Detta steg är något mer komplicerat eftersom fördelningen kan väljas relativt fritt. Vi vet från (27) att τ^2 alltid är icke-negativ men kan vara (nära) noll. Dessutom observeras ofta ett litet antal studier i i de flesta metaanalyser, vilket innebär att det är svårt att förvänta sig att det finns tillräckligt med information i datan för att dra slutsatser om mellan-studie-heterogenitet (Valentine, Pigott och Hannah R. Rothstein 2010, s. 31). Mot bakgrund av detta är det vanligt att välja *informativa* a priorifördelningar (Röver 2020, s. 9–12). Informativa a priorifördelningar, i motsats till icke-informativa eller svagt informativa fördelningar som inte starkt påverkar a posteriorifördelningen, kan minska osäkerheten i den a posteriorifördelning som erhålls. Informativa a priorifördelningar är ofta baserade på tidigare studier eller bedömningar från experter och kan användas för att inkorporera kända egenskaper hos parametern, såsom dess fördelning eller variationsvidd, i analysen. De kan också ta hänsyn till skillnader i stickprovsstorlek eller utförandet av studierna, vilka kan påverka osäkerheten i skattningarna.

En viktig aspekt vid användningen av informativa fördelningar är att säkerställa att den a priori-information som används är tillförlitlig och opartisk samt att den inte har alltför stor inverkan på analysen. Dessutom är det viktigt att välja en fördelning som är lämplig för den specifika forskningsfrågan och de data som analyseras (Schulz m. fl. 2021, s. 1).

Vid genomförandet av en metaanalys utan tidigare känd information kan det vara passande att undvika användningen av informativa fördelningar och istället hålla sig till svagt informativa sådana. En svagt informativ a priorifördelning som ofta rekommenderas för varianser liknande τ^2 är Half-Cauchyfördelningen (Röver 2020, s. 9). Half-Cauchyfördelningen är ett specialfall av en Cauchyfördelning, som endast definieras för ena halvan (den positiva sidan) av fördelningen. Fördel-ningen bestäms av två parametrar; platsparametern x_0 och skalparametern γ, där $x_0 = 0$ och $\gamma = 1$ ger en standardiserad Cauchyfördelning. Skalparametern γ styr spridningen av sannolikhetsfördelningen och påverkar hur mycket värdena av den stokastiska variabeln sprids ut från dess förväntade värde eller medelvärde. Half-Cauchyfördelningen betecknas med $\mathcal{HC}(x_0, \gamma)$.

Andra a priorifördelningar som har föreslagits i litteraturen inkluderar *Half-Normalfördelningen* $\mathcal{HN}(\mu_{HN}, \sigma_{HN})$ (svagt informativ) och den likformiga fördelningen $\mathcal{U}(a, b)$ (icke-informativ) (Klein och Kneib 2016, s. 1). Half-normalfördelningen är en normalfördelning med medelvärdet noll, men begränsad till att ha endast positiva värden. Skalningsparametern σ_{HN} representerar standardavvikelsen för den underliggande normalfördelningen. Om vi antar att en stokastisk variabel X följer en normalfördelning $X \sim \mathcal{N}(0, \sigma_N^2)$, kan vi uttrycka det som

$$|X| \sim \mathcal{HN}(\text{skala} = \sigma_N).$$

Den likformiga fördelningen är användbar som a priorifördelning i situationer där det saknas förkunskaper eller grund för att stödja något specifikt värde för τ. Denna fördelning har en konstant täthetsfunktion inom det definierade intervallet och är noll utanför. Fördelningen betecknas med $\mathcal{U}(a, b)$ där a och b representerar den nedre respektive övre gränsen för intervallet.

2.5.2 Likelihood

För att beräkna a posteriorifördelningen krävs, utöver en a priorifördelning, en likelihood-funktion. I fallet med den hierarkiska random effects-modellen baseras likelihood-funktionen på antagandena i (25) och beskriver en normalmodell med okänt medelvärde och varians. Enligt (Röver 2020, s. 13) är likelihood-funktionen definierad enligt följande:

$$p(\vec{Y} \mid \mu, \tau, \vec{\sigma}) = (2\pi)^{-\frac{k}{2}} \times \prod_{i=1}^{k} \frac{1}{\sqrt{\sigma_i^2 + \tau^2}} \exp\left(-\frac{1}{2}\frac{(Y_i - \mu)^2}{\sigma_i^2 + \tau^2}\right), \tag{28}$$

där $\vec{Y}$ är en vektor som innehåller observationerna Y_i och $\vec{\sigma}$ innehåller standardfelen σ_i för varje observation.

I en multivariat modell är det vanligtvis av intresse att studera den *marginella a posteriorifördelningen* (eng. *marginal posterior distribution*) för en specifik

parameter, i detta fall μ och τ. I följande stycken kommer vi härleda den *marginella likelihooden* av τ och använda den för att beräkna den marginella a posteriorifördelningen för μ och τ i kommande avsnitt.

Likelihood-funktionen är till en början en funktion av båda parametrarna μ och τ, se (28). Den marginella likelihooden för τ erhålls genom integration över μ, med hänsyn till den valda a priorifördelningen för μ, som vi specificerade antingen som likformig eller normal i avsnitt 2.5.1. Vi kommer nu att härleda den marginella likelihooden baserat på valet av a priorifördelning för μ. Följande härledningar och matematiska uttryck är hämtade direkt från artikeln *Bayesian Random-Effects Meta-Analysis Using the bayesmeta R Package* av Christian Röver 2020 (se avsnitt 2.3.-2.4.). För fullständiga härledningar hänvisas den intresserade läsaren till den ursprungliga artikeln.

Likformig a priorifördelning

Vid valet av en likformig a priorifördelning för μ, $p(\mu)$, kan vi härleda den marginella likelihooden

$$p(\vec{Y} \mid \tau, \vec{\sigma}) = \int p(\vec{Y} \mid \mu, \tau, \vec{\sigma}) p(\mu) \mathrm{d}\mu. \tag{29}$$

I (29) fås den marginella likelihooden genom att integrera likelihooden från (28) över alla möjliga värden av μ, vilket resulterar i marginalisering av μ. Slutresultatet är en funktion som endast är beroende av heterogeneiteten τ och vektorn av standardfel $\vec{\sigma}$.

Under REM utvecklas integralen till

$$\begin{aligned} &p(\vec{Y} \mid \tau, \vec{\sigma}) = \\ &(2\pi)^{-\frac{k-1}{2}} \times \prod_{i=1}^{k} \frac{1}{\sqrt{\sigma_i^2 + \tau^2}} \times \exp\left(-\frac{1}{2} \frac{(Y_i - \widehat{\mu}(\tau))^2}{\sigma_i^2 + \tau^2}\right) \times \frac{1}{\sqrt{\sum_{i=1}^{k} \frac{1}{\sigma_i^2 + \tau^2}}}, \end{aligned} \tag{30}$$

där $\widehat{\mu}(\tau)$ är det betingade a posteriori-medelvärdet av μ för en given heterogeneitet τ. Det betingade medelvärdet och standardavvikelsen ges i sin tur av

$$\begin{aligned} \widehat{\mu}(\tau) &= \mathrm{E}[\mu \mid \tau, \vec{Y}, \vec{\sigma}] = \frac{\sum_{i=1}^{k} \frac{Y_i}{\sigma_i^2 + \tau^2}}{\sum_{i=1}^{k} \frac{1}{\sigma_i^2 + \tau^2}}, \\ \widehat{\sigma}(\tau) &= \sqrt{\mathrm{Var}(\mu \mid \tau, \vec{Y}, \vec{\sigma})} = \sqrt{\frac{1}{\sum_{i=1}^{k} \frac{1}{\sigma_i^2 + \tau^2}}}. \end{aligned} \tag{31}$$

Normal a priorifördelning

Vid valet av en normal a priorifördelning för μ kan den marginella likelihooden härledas på liknande sätt som för den likformiga. Om vi antar att a priorifördelningen för μ är normal med medelvärde μ_p och en varians σ_p^2, kan vi

definiera $p(\mu)$ med hjälp av dess täthetsfunktion

$$p(\mu) = \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{p}}} \exp\left(-\frac{1}{2}\frac{(\mu-\mu_{\mathrm{p}})^2}{\sigma_{\mathrm{p}}^2}\right).$$

Den marginella likelihooden är då

$$\begin{aligned} p(\vec{Y} \mid \tau, \vec{\sigma}) &= \int p(\vec{Y} \mid \mu, \tau, \vec{\sigma}) p(\mu) \mathrm{d}\mu \\ &= (2\pi)^{-\frac{k+1}{2}} \times \frac{1}{\sqrt{\sigma_{\mathrm{p}}^2}} \times \prod_{i=1}^{k} \frac{1}{\sqrt{\sigma_i^2+\tau^2}} \\ &\times \int \exp\left(-\frac{1}{2}\left[\frac{(\mu-\mu_{\mathrm{p}})^2}{\sigma_{\mathrm{p}}^2} + \sum_{i=1}^{k}\frac{(Y_i-\mu)^2}{\sigma_i^2+\tau^2}\right]\right) \mathrm{d}\mu. \end{aligned} \tag{32}$$

Det betingade medelvärdet och standardavvikelsen ges i sin tur av

$$\begin{aligned} \widehat{\mu}(\tau) &= \frac{\frac{\mu_{\mathrm{p}}}{\sigma_{\mathrm{p}}^2} + \sum_{i=1}^{k}\frac{Y_i}{\sigma_i^2+\tau^2}}{\frac{1}{\sigma_{\mathrm{p}}^2} + \sum_{i=1}^{k}\frac{1}{\sigma_i^2+\tau^2}} \\ \widehat{\sigma}(\tau) &= \frac{1}{\sqrt{\frac{1}{\sigma_{\mathrm{p}}^2} + \sum_{i=1}^{k}\frac{1}{\sigma_i^2+\tau^2}}}. \end{aligned} \tag{33}$$

2.5.3 A posteriorifördelning för μ och τ

Så länge en likformig eller normal a priorifördelning används för effektstorleken μ, kommer den betingade a posteriorifördelningen för μ givet ett heterogenitetsvärde τ, betecknad som $p(\mu \mid \tau, \vec{Y}, \vec{\sigma})$, att vara normal med medelvärdet $\widehat{\mu}(\tau)$ och standardavvikelsen $\widehat{\sigma}$ enligt (31) eller (33) (Röver 2020, s. 15).

Efter att ha erhållit den marginella likelihooden $p(\vec{Y} \mid \tau, \vec{\sigma})$ enligt avsnitt 2.5.2, kan den marginella a posteriorítätheten av τ beräknas genom att multiplicera $p(\vec{Y} \mid \tau, \vec{\sigma})$ med a priorifördelningen för heterogeniteten $p(\tau)$, i enlighet med (23). Därmed erhålls följande ekvation för den marginella a posteriorifördelningen av τ:

$$p(\tau \mid \vec{Y}, \vec{\sigma}) \propto p(\vec{Y} \mid \tau, \vec{\sigma}) \times p(\tau). \tag{34}$$

Den simultana a posteriorifördelningen kan i sin tur uttryckas som en produkt av den marginella a posteriorifördelningen för τ och den betingade a posteriorifördelningen $p(\mu \mid \tau, \vec{Y}, \vec{\sigma})$ enligt

$$p(\mu, \tau \mid \vec{Y}, \vec{\sigma}) = p(\mu \mid \tau, \vec{Y}, \vec{\sigma}) \times p(\tau \mid \vec{Y}, \vec{\sigma}). \tag{35}$$

Genom att använda uttrycket i (35) kan den marginella a posteriorifördelningen för μ beräknas genom att integrera simultanfördelningen över alla möjliga värden

av τ. Detta innebär att vi marginaliserar bort heterogeniteten, vilket ger oss

$$p(\mu \mid \vec{Y}, \vec{\sigma}) = \int p(\mu, \tau \mid \vec{Y}, \vec{\sigma}) \mathrm{d}\tau = \int p(\mu \mid \tau, \vec{Y}, \vec{\sigma}) \times p(\tau \mid \vec{Y}, \vec{\sigma}) \mathrm{d}\tau \tag{36}$$

(Röver 2020, s. 16). Fördelningen i (36) är en kontinuerlig *blandningsfördelning* (eng. *mixture distribution*) där $p(\mu \mid \tau, \vec{Y}, \vec{\sigma})$ är "blandad" med $p(\tau \mid \vec{Y}, \vec{\sigma})$. En blandningsfördelning är en sannolikhetsfördelning som kombinerar två eller flera sannolikhetsfördelningar till en gemensam. (Röver och Friede 2017, s. 2).

En stokastisk variabel X anses ha en blandningsfördelning om dess fördelning beror på en annan variabel som i sig har en fördelning. Generellt sett leder hierarkiska modeller till blandningsfördelningar (Casella och Berger 2002, s. 165). Blandade fördelningar möjliggör för mer flexibilitet vid modellering av komplex data som inte enkelt kan beskrivas med en fördelning. Blandningsfördelningar används ofta för att modellera heterogenitet, men de kan också vara användbara som verktyg för att uppnå en god passning till data när standardfördelningar inte är tillräckliga (Seidel 2011, s. 1).

Den marginella a posteriorifördelningen för μ i ekvation (36), som är en normalblandad fördelning, kan approximeras. För att göra detta kan vi välja en uppsättning τ-värden och summera de betingade sannolikheterna för μ, givet dessa τ-värden, med "vikter" w_j. Vikterna beräknas med hänsyn till sannolikhetsfördelningen av τ. Följande approximation ges:

$$p(\mu \mid \vec{Y}, \vec{\sigma}) = \int p(\mu \mid \tau, \vec{Y}, \vec{\sigma}) \times p(\tau \mid \vec{Y}, \vec{\sigma}) \mathrm{d}\tau \approx \sum_j p(\mu \mid \tau_j, \vec{Y}, \vec{\sigma}) w_j \tag{37}$$

där den diskreta mängden τ_j väljs på ett lämpligt sätt och vikterna w_j, som uppfyller $\sum_j w_j = 1$, baseras på $p(\tau)$ (Röver 2020, s. 16).

För att approximera marginalfördelningen av en parameter i en blandningsmodell finns det flera tillgängliga metoder. Ett vanligt tillvägagångssätt är att använda numeriska integrationsmetoder, såsom Monte Carlo-integration eller Laplace approximation (Held och Sabanés Bové 2014, s. 247). I denna uppsats kommer vi att fokusera på DIRECT-metoden, som föreslagits av Röver och Friede (Röver och Friede 2017). Denna algoritm är speciellt utformad för problem med ett större antal variabler, vilket ofta är fallet i blandningsmodeller, och har visat sig vara effektiv och tillförlitlig när det gäller approximation av integraler (Röver och Friede 2017, s. 1).

2.5.4 DIRECT-metoden

I denna metaanalys ställs vi inför det vanligt förekommande problemet att härleda två parametrar, τ och μ, baserat på deras a priorifördelning. Från (37) kan $p(\tau \mid \vec{Y}, \vec{\sigma})$ härledas analytiskt och vi har att den betingade fördelningen $p(\mu \mid \tau, \vec{Y}, \vec{\sigma})$ är normal. Det primära intresset ligger vanligtvis i μ och genom att tillämpa DIRECT-metoden kan vi snabbt och exakt beräkna den marginella a posteriorifördelningen $p(\mu \mid \vec{Y}, \vec{\sigma})$. Denna metod baseras på *Kullback-Leibler divergens* och i detta avsnitt definieras alla relevanta begrepp. Följande

härledningar och matematiska uttryck är hämtade direkt från artikeln *Discrete Approximation of a Mixture Distribution via Restricted Divergence* av Christian Röver och Friede 2017 (se avsnitt 2.-4.3). För fullständiga härledningar hänvisas den intresserade läsaren till den ursprungliga artikeln.

Kullback-Leibler-divergens

Kullback-Leibler-divergensen (KL-divergensen) är ett mått på hur olika två sannolikhetsfördelningar är från varandra, genom att mäta mängden av information som går förlorad när man approximerar en fördelning med en annan (Bishop och Nasrabadi 2006, s. 77). Ju mindre KL-divergensen är, desto närmare är den approximativa fördelningen den sanna fördelningen. För två sannolikhetsfördelningar $p(\theta)$ och $q(\theta)$ definieras KL-divergensen enligt

$$\begin{aligned} \mathcal{D}_{\mathrm{KL}}(p(\theta)\|q(\theta)) &= \int_{\Theta} \log\left(\frac{p(\theta)}{q(\theta)}\right) p(\theta) d\theta \\ &= \mathrm{E}\left[\log\left(\frac{p(\theta)}{q(\theta)}\right)\right] \end{aligned} \tag{38}$$

(Röver och Friede 2017, s. 1; Held och Sabanés Bové 2014, s. 329). Divergensen är alltid positiv och är noll när de två fördelningarna är identiska ($p = q$). Generellt gäller att $\mathcal{D}_{\mathrm{KL}}$ inte är symmetrisk, vilket innebär att

$$\mathcal{D}_{\mathrm{KL}}(p(\theta)\|q(\theta)) \neq \mathcal{D}_{\mathrm{KL}}(q(\theta)\|p(\theta)).$$

För att hantera detta definieras istället den *symmetriska KL-divergensen* enligt

$$\mathcal{D}_{\mathrm{s}}(p(\theta)\|q(\theta)) = \mathcal{D}_{\mathrm{KL}}(p(\theta)\|q(\theta)) + \mathcal{D}_{\mathrm{KL}}(q(\theta)\|p(\theta)) \tag{39}$$

(Röver och Friede 2017, s. 1). I följande text kommer vi fokusera på den symmetriska KL-divergensen.

Binning

Nu vill vi definiera en metod för att approximera en blandningsfördelning och inför i följande text vissa variabler och notationer för att underlätta förståelsen. Vi antar att en stokastisk variabel Z följer en täthetsfunktion $p(z \mid x)$, som beror på en stokastisk variabel X med täthet $p(x)$. Då är den marginella fördelningen för Z, $p(z)$, en blandningsfördelning [1]. För att beräkna marginalfördelningen för Z integrerar vi simultanfördelningen $p(x, z)$ över x, vilket ger oss

$$p(z) = \int p(x, z) dx = \int p(z \mid x) p(x) dx.$$

Blandningsfördelningen $p(z)$ är kontinuerlig och för att kunna approximera denna kan det vara lämpligt att använda diskreta fördelningar. För att göra detta definierar vi en indelning, en så kallad *binning*, av definitionsmängden för X

[1] I vårt fall motsvarar $p(z)$ a posteriorifördelningen $p(\mu \mid \vec{Y}, \vec{\sigma})$.

(Röver och Friede 2017, s. 2). Vi låter $\left\{x_{(1)}, x_{(2)}, \ldots, x_{(m-1)}\right\} \subset \mathbb{R}$ vara en mängd av gränser för m stycken indelningar, där vi antar att $x_{(1)} < x_{(2)} < \cdots < x_{(m-1)}$. Den diskreta mängden $\{\mathcal{X}_i\}_{i=1,\ldots,m}$ med m stycken indelningar definieras enligt

$$\mathcal{X}_i = \begin{cases} \left\{x : x \leq x_{(1)}\right\} & \text{om } i = 1 \\ \left\{x : x_{(i-1)} < x \leq x_{(i)}\right\} & \text{om } 1 < i < m \\ \left\{x : x_{(m-1)} < x\right\} & \text{om } i = m. \end{cases} \tag{40}$$

Notationen {x: villkor} avser mängden av alla x-värden som uppfyller det givna villkoret. Således definierar $\mathcal{X}_i$ en mängd av x-värden som kategoriseras enligt en specifik diskret indelning, som i sin tur är definierad av gränserna $\left\{x_{(1)}, x_{(2)}, \ldots, x_{(m-1)}\right\}$. Den första indelningen innehåller alla X-värden som är mindre än eller lika med $x_{(1)}$, den andra innehåller alla X-värden som är större än $x_{(1)}$ men mindre än eller lika med $x_{(2)}$, och så vidare. Den sista indelningen inkluderar alla värden X-värden som är större än $x_{(m-1)}$. För varje indelning finns en sannolikhet π_i (med avseende på $p(x)$) associerad, vilken ges av

$$\pi_i = \mathrm{P}\left(x_{(i-1)} < x \leq x_{(i)}\right) = \mathrm{P}\left(x \in \mathcal{X}_i\right). \tag{41}$$

Dessutom har vi $\{\tilde{x}_1, \ldots, \tilde{x}_m\}$ med $\tilde{x}_i \in \mathcal{X}_i$ som definierar en mängd med k stycken *referenspunkter*, en för varje indelning. Hur dessa referenspunkter väljs diskuteras i Algortim 1.

Utöver den angivna tätheten $p(x, z)$ introducerar vi en annan sannolikhetsfördelning $q(x, z)$ som har samma marginaltäthet

$$q(x) = p(x),$$

och vars betingade sannolikhetsfördelning ges av

$$q(z \mid x) = p\left(z \mid x = \tilde{x}_i\right) \qquad \text{för } x \in \mathcal{X}_i \tag{42}$$

(Röver och Friede 2017, s. 2). Vi har alltså att q är lik p, men här betingas sannolikhetsfördelningen på motsvarande referensvärde $\tilde{x}_i$ för den tillhörande indelningen m, istället för att betingas på det exakta x-värdet. Precis som med simultanfördelningen $p(x, z)$ definieras $q(x, z) = q(z \mid xq(x))$ och marginaltätheten för Z är $q(z) = \int q(z \mid x)q(x)dx$. Vi kan alltså betrakta denna indelningsprocess som ett sätt att diskretisera tätheten för $p(x)$, fördelningen för den underliggande variabeln, samtidigt som vi håller den betingade fördelningen oförändrad. Den diskretiserade versionen har referenspunkterna $\{\tilde{x}_1, \ldots, \tilde{x}_m\}$ som sin definitionsmängd och sannolikheterna $\{\pi_1, \ldots, \pi_m\}$ definierar dess sannolikhetsmassfunktion.

Genom att kunna skriva om simultanfördelningen av X och Z i diskreta termer kan marginalfördelningen $q(z)$ skrivas som en summa istället för en integral (Röver och Friede 2017, s. 3). Uttrycket för marginaltätheten blir då

$$q(z) = \sum_{i=1}^{m} \pi_i p\left(z \mid \tilde{x}_i\right), \tag{43}$$

vilket möjliggör för enklare approximation.

Vidare definieras den maximerade symmetriska KL-divergensen enligt

$$d_i = \max_{x \in \mathcal{X}_i} \{\mathcal{D}_\mathrm{s}\left(p(z \mid x) \| p\left(z \mid \tilde{x}_i\right)\right)\} = \max_{x \in \mathcal{X}_i} \{\mathcal{D}_\mathrm{s}(p(z \mid x) \| q(z \mid x))\}, \tag{44}$$

vilket är en maxfunktion som ger den största divergensen mellan sannolikhetsfördelningarna $p(z \mid x)$ för punkter inom den i:e indelningen och den motsvarande i:e referenspunkten. Kedjeregeln för Kullback-Leibler-divergensen säger att

$$\mathcal{D}_\mathrm{KL}(p(x,z) \| q(x,z)) = \mathcal{D}_\mathrm{KL}(p(x) \| q(x)) + \mathrm{E}\left[\mathcal{D}_\mathrm{KL}(p(z \mid x) \| q(z \mid x))\right], \tag{45}$$

det vill säga att KL-divergensen kan skrivas som summan av två termer: KL-divergensen mellan $p(x)$ och den diskretiserade versionen $q(x)$, samt väntevärdet av KL-divergensen mellan den betingade fördelningen $p(z \mid x)$ och motsvarande diskretiserade versionen $q(z \mid x)$ (Röver och Friede 2017, s. 3; Cover och Thomas 1991, s. 24–25). På motsvarande sätt gäller kedjeregeln för den symmetriska divergensen enligt

$$\mathcal{D}_s(p(x,z) \| q(x,z)) = \mathcal{D}_s(p(x) \| q(x)) + \mathrm{E}\left[\mathcal{D}_s(p(z \mid x) \| q(z \mid x))\right]. \tag{46}$$

Då vi definierat att $p(x) = q(x)$ följer att KL-divergensen mellan dessa fördelningar är noll, det vill säga $\mathcal{D}_\mathrm{KL}(p(x) \| q(x)) = \mathcal{D}_\mathrm{KL}(q(x) \| p(x)) = 0$. Av detta följer att (46) kan skrivas om som

$$\mathcal{D}_\mathrm{s}(p(x,z) \| q(x,z)) = \mathrm{E}\left[\mathcal{D}_s(p(z \mid x) \| q(z \mid x))\right]. \tag{47}$$

Vi är främst intresserade av approximationen av $p(z)$, med hjälp av den förenklade fördelningen $q(z)$. Genom användning av kedjeregeln härleds följande uttryck:

$$\begin{aligned} &\mathcal{D}_\mathrm{s}(p(z) \| q(z)) = \mathcal{D}_\mathrm{s}(p(x,z) \| q(x,z)) - \mathrm{E}\left[\mathcal{D}_s(p(x \mid z) \| q(x \mid z))\right] \\ &\leq \mathcal{D}_\mathrm{s}(p(x,z) \| q(x,z)) = \mathrm{E}\left[\mathcal{D}_s(p(z \mid x) \| q(z \mid x))\right] \\ &\leq \sum_i \pi_i d_i \leq \max_i d_i =: \delta \end{aligned} \tag{48}$$

(Röver och Friede 2017, s. 3).

Uttrycket i (48) innebär att om vi sätter en övre gräns, eller begränsning, på divergensen mellan $p(z|x)$ och $q(z|x)$ för varje individuell indelning i, där denna gräns är mindre än eller lika med δ, kan vi också begränsa divergensen mellan de exakta och approximativa marginalfördelningarna $p(z)$ och $q(z)$. Genom att kontrollera divergensen inom varje indelning kan vi säkerställa att skillnaden mellan den exakta och approximativa fördelningen som helhet är begränsad (Röver och Friede 2017, s. 3). Frågan som uppstår är om vi kan skapa en form av rutnätsapproximation (eng. *grid approximation*) med avseende på denna maximala divergens δ och i så fall hur detta kan göras.

Vi inleder processen genom att fastställa en referenspunkt x i definitionsmängden för den underliggande funktionen. Därefter identifierar vi närliggande punkter kring x, där divergensen mellan referenspunkten och dessa punkter är mindre än δ, och låter denna grupp av punkter definiera en indelning. Genom att tillämpa denna indelning kan vi täcka hela utfallsrummet för Z. Denna metod, känd som DIRECT (divergence restricting conditional tesselation)-metoden, syftar till att täcka ufallsrummet samtidigt som divergensen begränsas (Röver och Friede 2017, s. 3).

I vissa fall är det inte möjligt att uppnå en uppdelning så att alla indelningar har ändlig divergens. För att komma runt detta problem kan vi bortse från en liten del av utfallsrummet, som associeras med en godtycklig sannolikhet ϵ, och genomföra indelningen i resten av utfallsrummet. Genom att göra detta exkluderar vi en bråkdel av värden som förekommer mycket sällan, och fokuserar istället på resten av fördelningen. Dessa värden som inte passar in i någon av indelningarna utgör vanligtvis inte ett betydande problem. Det beror på att de associeras med en godtyckligt liten sannolikhet ϵ och behandlas istället som feltermer i approximationen.

Algoritmen

Målet är att konstruera en indelning som ger en noggrann diskret approximation av marginaltätheten genom användning av KL-divergenser. Idén är att dela upp värdena i definitionsmängden för den underliggande fördelningen i m stycken mindre indelningar, samtidigt som vi säkerställer att divergensen inom varje indelning m är begränsad. Om det uppstår extrema värden, såsom långt ut i fördelningens svansar, bör dessa ignoreras (Röver och Friede 2017, s. 3). Vidare antar vi att divergensen mellan ett godtyckligt par av punkter (x_1, x_2) i utfallsrummet är *Lipschitz-kontinuerlig*, vilket innebär att det finns en gräns för hur stor skillnad det får finnas mellan de två punkterna. Denna gräns antas gälla åtminstone inom ett visst område av utfallsrummet, specificerat som intervallet $[\tilde{x}_1, \tilde{x}_k]$. Algoritmen förutsätter också att sannolikheten för att en observation befinner sig utanför detta intervall är mindre än eller lika med ϵ, det vill säga $\mathrm{P}\left(X \notin [\tilde{x}_1, \tilde{x}_k]\right) \leq \epsilon$ (Röver och Friede 2017, s. 3). Röver beskriver algoritmen med sex steg, vilka beskrivs i Algoritm 1.

Algoritmen strävar efter att minimera KL-divergensen mellan intilliggande indelningar samtidigt som kravet på en minimivikt π_i för varje indelning i uppfylls. Referenspunkterna $\tilde{x}_i$ och motsvarande vikter π_i möjliggör nu för att definiera en approximation q enligt ekvation (43). Här är det viktigt att försäkra sig om att villkoret

$$\sum_i \pi_i d_i \leq \max_i d_i =: \delta$$

är uppfyllt. Möjliga gräns- eller singularitetsproblem undviks genom att ignorera försumbara delar av utfallsrummet via specificering av ϵ (Röver och Friede 2017, s. 4). Att divergensen är Lipschitz-kontinuerlig säkerställer att den relevanta mängden kan täckas av ett ändligt antal indelningar.

Algoritm 1 DIRECT-metoden

1: Ange en maximal KL-divergens $\delta > 0$, en liten sannolikhet $0 \leq \epsilon \ll 1$, och en startreferenspunkt $\tilde{x}_1$. Förnuftiga värden för $\tilde{x}_1$ kan till exempel vara det minsta möjliga värdet, något värde med $\mathrm{P}\left(X \leq \tilde{x}1\right) < \epsilon$ eller $\frac{\epsilon}{2}$-kvantilen. Definiera $\epsilon_1 := \mathrm{P}\left(X \leq \tilde{X}_1\right) \geq 0$. Sätt $i = 1$.

2: Sätt $x^\star = \tilde{x}_1$. Av det följer att $\mathcal{D}s\left(p\left(z \mid \tilde{x}_1\right) | p\left(z \mid x^\star\right)\right) = 0$. Öka nu $x^\star$ så långt som möjligt samtidigt som det säkerställs att $\mathcal{D}\mathrm{s}\left(p\left(z \mid \tilde{x}1\right) | p\left(z \mid x^\star\right)\right) \leq \delta$. Använd sedan denna punkt som den första gränsvärdet för bin: $x(1) = x^\star$. Beräkna $\pi_1 = \mathrm{P}\left(x < x_{(1)}\right)$. Sätt $i = i + 1$.

3: Öka värdet på $x^\star$ tills $\mathcal{D}_s\left(p\left(z \mid x_{(i-1)}\right) \| p\left(z \mid x^\star\right)\right) = \delta$. Använd sedan denna punkt som nästa referenspunkt: $\tilde{x}_i = x^\star$.

4: Öka $x^\star$ igen tills $\mathcal{D}_\mathrm{s}\left(p\left(y \mid \tilde{x}_i\right) \| p\left(z \mid x^\star\right)\right) = \delta$. Använd denna punkt som gränsvärde för indelningen: $x_{(i)} = x^\star$

5: Beräkna indelningens vikt $\pi_i = \mathrm{P}\left(x_{(i-1)} < X \leq x_{(i)}\right)$.

6: Om $\mathrm{P}\left(X > x_{(i)}\right) > (\epsilon - \epsilon_1)$, sätt $i = i + 1$ och gå tillbaka till steg 3. Annars avsluta algoritmen.

2.6 Publiceringsbias

Samtidigt som metaanalyser ger en matematiskt korrekt och noggrann syntes av flera studier, finns det alltid en risk att studierna som är inkluderade i analysen inte är tillräckligt representativa för studierna inom området (Hannah R Rothstein, Sutton och Borenstein 2005, s. 15). Det största hotet är känt som *publiceringsbias* och är ett genomgående problem inom vetenskaplig forskning. Detta fenomen kan ha allvarliga konsekvenser för metaanalyser och andra relaterade metoder för systematisk litteraturöversikt. Publiceringsbias uppstår när studier med positiva eller statistiskt signifikanta resultat är mer benägna att publiceras, medan studier med icke-signifikanta eller negativa resultat (resultat som går åt motsatt håll än förväntat) har mindre sannolikhet att bli publicerade (Rosenthal 1979, s. 1). Detta kan troligtvis förklaras av att forskare är mindre benägna att rapportera negativa resultat, samtidigt som det finns en tendens hos vetenskapliga tidskrifter att vara mindre villiga att publicera studier med sådana resultat (Olson m. fl. 2002, s. 1; Thornton och Lee 2000, s. 9). Andra faktorer som kan bidra till partiskhet i valet av studier, som faller under termen publiceringsbias, inkluderar följande (Borenstein m. fl. 2009, s. 280):

- Språk. Artiklar och tidskrifter som är skrivna på engelska är mer populära och eftersökta, vilket leder till att urvalet av studier som väljs till metaanalysen ofta är överrepresenterat med statistiskt signifikanta studier.

- Tillgänglighet och kostnad. Det kan förekomma partiskhet vid val av studier baserat på tillgänglighet snarare än deras representativitet inom området. Kostnaden för att få tillgång till artikeln spelar också roll, där gratis eller lågprisartiklar oftare inkluderas jämfört med motsvarande alternativ.

- Duplicering. Studier med statistiskt signifikanta resultat har större benägenhet att publiceras mer än en gång, vilket ökar sannolikheten att de upptäcks och inkluderas i en metaanalys. Det är inte alltid uppenbart att flera publikationer kommer från en enda studie eller att en grupp studiedeltagare kan inkluderas i en analys mer än en gång. Inkludering av duplicerade data kan därmed leda till överskattning av den beräknade effekten.

- Citering. Studier med statistiskt signifikanta resultat har större sannolikhet att bli citerade av andra forskare och därigenom lättare att identifiera.

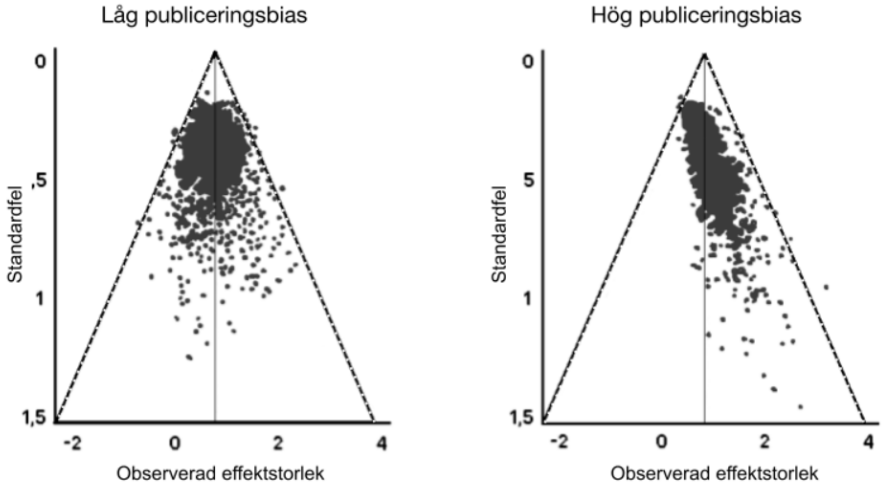
Denna selektiva publicering och sätt att välja studier på utgör en betydande problematik. Om en metaanalys endast inkluderar publicerade studier (under antagandet att detta problem existerar i området) kan det leda till en överestimering av den verkliga effektstorleken, och en felaktig representation av det övergripande forskningsmaterialet (Hannah R Rothstein, Sutton och Borenstein 2005, s. 26). I de mest allvarliga fallen kan felaktiga slutsatser dras, till exempel att en behandling anses vara effektiv trots att den inte är det eller att den kan ha negativa effekter.

För att ta itu med publiceringsbias i metaanalyser har flera metoder utvecklats för att mäta dess närvaro och omfattning. Ett vanligt tillvägagångssätt är att undersöka s.k. *funnelplottar* (eng. *funnel plots*). En funnelplott är bland annat ett grafiskt hjälpmedel för att undersöka partiskhet, där de observerade effektstorlekarna plottas mot antingen deras standardfel eller antalet deltagare (Egger m. fl. 1997, s. 1). I figur 1 presenteras två funnelplottar, där den vänstra visar tecken på låg publiceringsbias och den högra på hög publiceringsbias. Den streckade linjen i mitten av de två funnelplottarna motsvarar den genomsnittliga effektstorleken $\hat{\mu}$. De yttre diagonala linjerna representerar den triangulära regionen inom vilken 95% av studierna förväntas ligga, i frånvaro av både publiceringsbias och heterogenitet, vilket motsvarar $[\hat{\mu} \pm (1.96 \times \text{standardfel})]$ (Sterne och Egger 2001, s. 2; Hannah R Rothstein, Sutton och Borenstein 2005, s. 82). Större studier befinner sig främst högre upp i figuren och tenderar att samlas nära den genomsnittliga effektstorleken. Nedre delen av figuren visar på en större spridning, eftersom vi kan förvänta oss att mindre studier har en större variation i de skattade värdena.

I avsaknad av publiceringsbias bör plotten vara symmetrisk, med färre studier som visar större variation kring skattningen av effektstorleken på grund av provtagningsfel. Om det däremot finns publiceringsbias (till exempel genom att mindre studier som inte visar några statistiskt signifikanta effekter förblir opublicerade), kommer funnelplotten att uppvisa asymmetri med ett "tomrum" i något av de nedre hörnen.

Att bedöma asymmetrin i en funnelplott är ofta subjektivt och det finns mer formella sätt för att testa publikationsbias som bygger på funnelplotten. Ett av dessa sätt är *trim and fill*-metoden, som inte bara testar för bias utan också korrigerar för det vid beräkning av den genomsnittliga effektstorleken och standardfelet (Borenstein m. fl. 2009, s. 286). Metoden bygger på en iterativ process som består av två steg: trim och fill.

Figur 1: Publiceringsbias. Den vänstra grafen visar en metaanalys med låg förekomst av publiceringsbias, medan den högra grafen visar en metaanalys med hög förekomst av publiceringsbias.



I trim-stadiet tar metoden temporärt bort alla studier som bidrar till asymmetrin i funnelplotten. Därefter beräknas en ny effektstorleksskattning, i varje iteration, som inte påverkas av asymmetrin. Medan denna trimning ger en justerad effektstorlek, leder den även till minskad variation av effekterna, vilket i sin tur resulterar i ett smalt konfidensintervall. För att hantera detta återinförs effektstorlekar i fill-stadiet, vilket återigen skapar asymmetri. Effektstorlekar som nu saknas"imputeras i fill-stadiet för att uppnå symmetri i funnelplotten. Det är viktigt att notera att denna fill"inte påverkar själva punktskattningen, utan syftar istället till att korrigera variansen. Denna tvåstegsprocess upprepas tills ett givet konvergenskriterium är uppfyllt, det vill säga när dessa två steg inte längre förändrar skattningen av den genomsnittliga effektstorleken.

3 Data

Detta arbete syftar till att genomföra en metaanalys för att bedöma validiteten av anställningsintervjuer som en prediktor för faktisk arbetsprestation. Datamaterialet som används hämtas från **dat.mcdaniel1994**-datasetet, som finns tillgängligt i R-paketet `metafor`. Dataset inkluderar $k = 160$ studier, med totalt 25 244 deltagare, som undersöker sambandet mellan prestation vid anställningsintervjuer och faktisk arbetsprestation. McDaniel m.fl. (1994) undersökte detta samband i en metaanalys baserat på dessa 160 studier, där målet var att undersöka om anställningsintervjuer är ett effektivt verktyg för att välja medarbetare som kommer att prestera väl i arbetsmiljön. En specifik hypotes var att det finns en linjär relation mellan intervjuresultat och jobbprestation.

Effektstorleken som användes för att kvantifiera detta samband var korrelationskoefficienten, representerad som r.

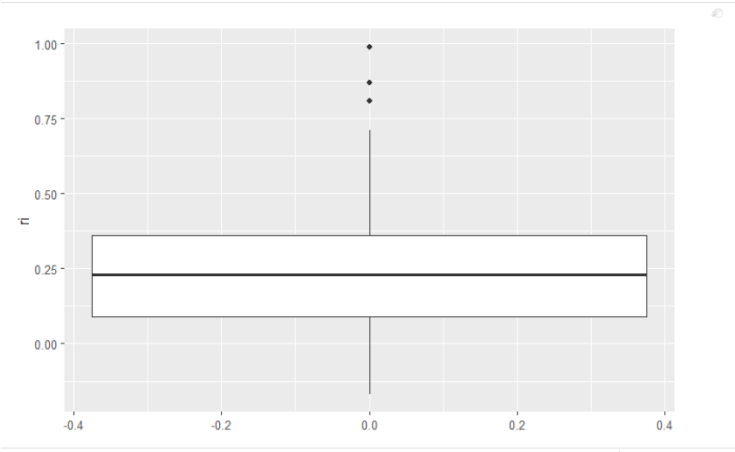
3.1 Beskrivning av variabler

Datamaterialet från **dat.mcdaniel1994** ger information om följande variabler:

1. **study**: Index för varje individuell studie.
2. **ni**: Stickprovsstorleken i varje studie.
3. **ri**: Den observerade korrelationskoefficienten mellan intervjuprestationen och den faktiska arbetsprestationen.
4. **type**: Typen av anställningsintervju som användes i varje studie (situationsbaserad, jobbrelaterad eller psykologisk).
5. **struct**: Strukturen för anställningsintervjun som användes i varje studie (strukturerad eller ostrukturerad).

Spridningen av korrelationerna **ri** illustreras med ett lådagram i figur 2. De rapporterade korrelationskoefficienterna **ri** varierar mellan -0.17 och 0.99, med ett medelvärde på 0.24 och en standardavvikelse på 0.21. Stickprovsstorleken **ni** varierar mellan 6 och 4195 deltagare, med ett genomsnitt på 157.

Figur 2: Lådagram som illustrerar spridningen hos de observerade korrelationskoefficienterna **ri**.



Variablerna intervjutyp (**type**) och intervjustruktur (**struct**) ger ytterligare information för att utforska potentiella moderatorer av sambandet mellan prestation vid intervju och arbetsprestation. Situationsbaserade intervjuer fokuserar på en individs förmåga att hantera givna situationer, arbetsrelaterade intervjuer

bedömer tidigare erfarenheter samt information relaterad till arbetsuppgifter, medan psykologiska intervjuer bedömer personliga egenskaper. Vidare är strukturerade intervjuer kännetecknade av specifika frågor och en mer formell natur, medan ostrukturerade intervjuer samlar information på ett mindre systematiskt och mer spontant sätt.

Mcdaniel1994-datasetet utgör en betydande informationskälla för att genomföra en metaanalys av anställningsintervjuers validitet och utforska potentiella orsaker till dessa samband.

3.2 Transformation av data

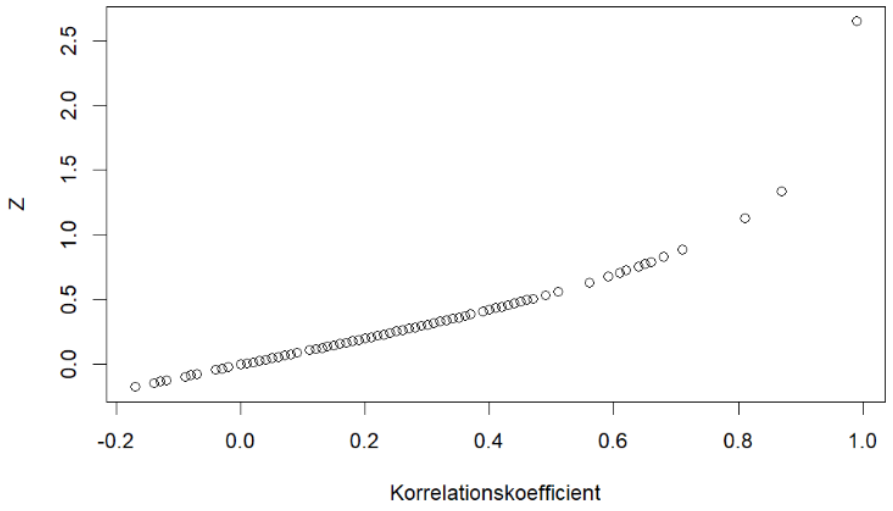
Vid utförandet av en metaanalys där effektstorleken mäts i form av korrelationskoefficienter är det nödvändigt att transformera koefficienterna med hjälp av Fishers z-transformation, se avsnitt 2.1.1. Transformationen (se (2 och 3) tillämpas på **ri**, vilket resulterar i två nya variabler:

1. **yi**: Den observerade korrelationskoefficienten **ri** i Fishers z-skala.

2. **vi**: Variansen för **yi**.

Resultatet av transformationen illustreras i figur 3, där korrelationskoefficienterna återfinns på x-axeln och de transformerade värdena på y-axeln. Figuren visar att efter z-transformen vägs korrelationsvärden nära 1 tyngre, d.v.s. de har större z-värden. Denna viktningsstrategi motiveras av det faktum att det är betydligt mer osannolikt att erhålla en korrelationskoefficient på 0.9 i en studie, jämfört med en på 0.2. Följaktligen bör korrelationer med högre magnitud tilldelas större vikt. Detta är särskilt relevant när vi vill testa den genomsnittliga effekten μ mot 0. Då är det viktigt att ha i åtanke att vissa korrelationsvärden var nära 1. Om dessa värden tros vara resultatet av slumpmässiga variationer kan detta bero på att standardfelet är stort, vilket i sin tur indikerar en liten stickprovsstorlek. Enligt (4) är standardfelet för en korrelation omvänt proportionellt mot stickprovsstorleken. Av den anledningen kan vi tilldela mindre vikt åt studier som visar höga korrelationer men har mycket små stickprovsstorlekar, då höga korrelationer i sådana fall ibland kan uppstå av ren slump!

Figur 3: Observerad korrelation mellan intervjubedömning och arbetsprestation.



4 Metaanalys

Denna studie syftar till att replikera McDaniels (1994) metaanalys och utöka den genom att utföra två separata metaanalyser med olika statistiska metoder. Den första metaanalysen kommer att använda frekventistiska metoder med hjälp av R-paketet `metafor`, medan den andra kommer att använda Bayesianska metoder med paketet `bayesmeta` som implementerar DIRECT-algoritmen.

4.1 Frekventistisk

En frekventistisk metaanalys med random effects-modellen utfördes för att uppskatta den totala effektstorleken av arbetsintervjuebedömningar för arbetsprestation. Analysen inkluderade alla 160 studier och testade nollhypotesen

$$H_0 : \mu = 0.$$

Den uppskattade totala effektstorleken, enligt metoden beskriven i avsnitt 2.4, var $\hat{\mu} = 0.2368$ med ett 95% konfidensintervall på (0.2046, 0.2690) och $p < 0.0001$. För enklare tolkning omvandlades skattningen av μ tillbaka till korrelationsskala och gav då en korrelation på $\hat{\rho} = 0.2325$ (95% KI:0.2018 till 0.2627). Med det låga p-värdet förkastades nollhypotesen och resultaten visade att korrelationen mellan arbetsintervjubedömningar och arbetsprestation var positiv, men svag.

Mängden heterogeneitet för effektstorleken uppskattades vara $\hat{\tau}^2 = 0.0268$ med ett 95% konfidensintervall (0.0240, 0.0502). Värdet på τ var i sin tur 0.1636,

vilket indikerar att de sanna effektstorlekarna har en uppskattad standardavvikelse på ca 0.16, uttryckt i Fishers z-skalan. För att testa om det fanns signifikant heterogeneitet utfördes Cochrans Q-test, och resultatet visade att $Q = 789.73 > \chi^2_{0.05}(159)$ vilket var betydligt högre än förväntat baserat på $k - 1 = 159$ frihetsgrader. Därmed var testet signifikant ($p < 0.001$), men det bör noteras att bedömningen inte enbart bör baseras på Q-testet, som tidigare nämnts. Vidare fann vi att $I^2 = 79.87\%$, vilket indikerar att majoriteten av variationen mellan effektstorlekarna orsakades av heterogenitet mellan studierna snarare än stickprovsfel. Enligt Higgins och Thompsons tumregler (se avsnitt 2.2.2.1) karakteriseras mängden heterogenitet som hög. Resultaten tyder på att det finns signifikant heterogenitet mellan effektstorlekarna i de inkluderade studierna. Slutligen noterades att konfidensintervallet för $\hat{\tau}^2$ inte innehöll noll, vilket överensstämmer med teorin som säger att $\tau > 0$.

4.1.1 Undergruppsanalys

I analysen undersöktes även om intervjuformen påverkade effektstorleken μ. Datamängden delades upp baserat på typ av intervju och intervjuens struktur, enligt beskrivningarna i avsnitt 3.1. Totalt klassificerades 157 studier som antingen arbetsrelaterade, situationsbaserade eller psykologibaserade, medan tre studier saknade klassificering. Vad gäller intervjuens struktur, var 145 studier antingen strukturerade eller ostrukturerade, medan 15 studier saknade information om strukturen.

En metaanalys utfördes på dessa nya datagrupperingar, där varje intervjuform fick en ny skattning av μ och τ^2. Skattningarna från dessa analyser, tillsammans med de tidigare utförda av McDaniel m.fl., sammanställdes i tabell 1, där $\hat{\mu}$ omvandlades till korrelationen $\hat{\rho}$.

Tabell 1: Frekventistiska analysresultat av korrelation mellan intervjuer och arbetsprestation, enligt intervjutyp och struktur.

Intervjuform	Antal studier, k	$\hat{\rho}$ (McDaniel, 1994)	$\hat{\rho}$	$\hat{\tau}^2$
Arbetsrelaterad	127	0.28	0.2436	0.0334
Situationsbaserad	16	0.35	0.2852	0.0085
Psykologibaserad	14	0.20	0.1495	0.0015
Strukturerad	106	0.31	0.2659	0.0389
Ostrukturerad	39	0.23	0.1862	0.0093

Resultaten av de nyligen genomförda analyserna visade att $\hat{\rho}$ antog värden mellan ungefär 0.15 och 0.29 för de olika intervjuformerna. Mängden heterogenitet, uttryckt som τ^2, var signifikant högre i analyserna med ett större antal studier, vilket överensstämmer med antagandet om en random effects-modell. Samtliga fem metaanalyser visade att $p < 0.001$ och nollhypotesen kunde därav förkastas.

Tabell 1 visar att de nya skattningarna av korrelationen avviker något från de gjorda av McDaniel. En möjlig anledning till dessa små skillnader kan vara att olika formler användes för beräkningen.

4.2 Bayesiansk

För att genomföra en Bayesiansk random effects metaanalys var det nödvändigt att bestämma a priorifördelningar för både μ och τ. Vi genomförde empiriska undersökningar av olika kombinationer av a priorifördelningar för μ och τ, där både fördelningsparametrar och antalet studier varierades. Följande a priorifördelningar valdes för μ:

1. $\mu \sim \mathcal{N}(0, 0.5)$
2. $\mu \sim \mathcal{U}(-\infty, \infty)$.

Valet av en normalfördelning med medelvärde 0 och standard avvikelse 0.5 är ett neutralt a priori-antagande, då det saknades några förkunskaper eller starka teoretiska förväntningar om värdet på μ. Med dessa parametrar infördes inga starka antaganden eller gissningar på a priorifördelningen, utan vi tillät datan att "tala för sig själv". En standardavvikelse på 0.5 innebär att vi förväntar oss att de flesta korrelationskoefficienterna kommer att vara inom intervallet -0,5 och 0,5, vilket är rimligt för korrelationsdata. Förutom normalfördelningen valdes även en icke-informativ likformig fördelning för att tillåta en bredare variation av möjliga värden för μ.

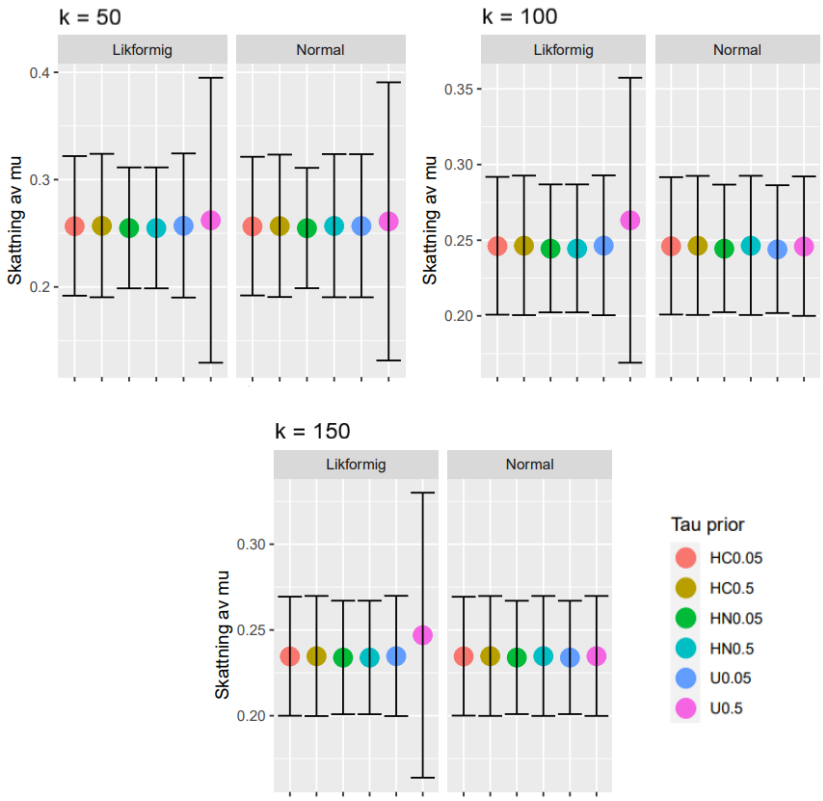
Följande a priorifördelningar valdes för τ:

1. $\tau \sim \mathcal{HN}(0, \beta)$
2. $\tau \sim \mathcal{HC}(0, \beta)$
3. $\tau \sim \mathcal{U}(\varphi, 1)$

där $\beta, \varphi \in \{0.05, 0.5\}$. I metaanalyser som behandlar data i en standardiserad skala, såsom Fishers z, är 0.5 ett vanligt värde på a posteriorimedianen för τ. Det beror på att medianvärden över 0.5 anses vara mindre troliga (Williams, Rast och Bürkner 2018, s. 4). Av denna anledning har 0.5 valts som parametervärde för τ-fördelningarna, medan värdet 0.05 används för att undersöka om det har en betydande inverkan på skattningarna om parametern är mycket mindre. Efter att a priorifördelningarna valts, genererades alla möjliga kombinationer av μ- och τ-fördelningar, vilka var totalt 12 kombinationer. Dessa användes sedan för att utföra metaanalyser med $k \in 50, 100, 150$ antal studier, vilket ger 12 olika analyser per analysstorlek k. Här valdes $k \in 50, 100, 150$ för att undersöka om resultaten är robusta över ett brett spektrum av studieantal, samtidigt som antalet studier är tillräckligt stort för att uppnå en god precision i skattningarna.

I figur 4 visas den skattade $\hat{\mu}$ för alla 36 analyser. Varje analys representeras av antingen en Likformig"eller Normalfördelning för μ, och färgerna indikerar fördelningen för τ. De presenterade intervallen utgör 95% kredibilitetsintervall för de skattade värdena.

Figur 4: Skattningar av μ för olika kombinationer av a priorifördelningar för μ och τ, vid analysstorlekarna $k \in 50, 100, 150$. Rubrikerna Likformigöch Normalrepresenterar de valda a priorifördelningarna för μ, medan de färgade cirklarna indikerar fördelningen för τ.



Resultaten visade att valet av a priori-fördelningar inte hade stor påverkan på den slutgiltiga skattningen. Värdet för $\hat{\mu}$ var i stort sett likvärdigt för de olika analysstorlekarna (k), med endast små decimalavvikelser. Dock observerades en något högre uppskattad effektstorlek för en mindre analysstorlek ($k = 50$), och en lägre skattning för en större analysstorlek ($k = 150$). Detta behöver inte nödvändigtvis bero på en faktisk skillnad i antalet studier k som analyseras, utan snarare på att studierna som ingår i varje analys (för $k = 50, 100$ och 150) kan skilja sig i sina stickprovsstorlekar och resultat. Med andra ord kan studierna som ingår i $k = 50$-analysen i genomsnitt ha haft mindre stickprovsstorlekar jämfört med de som ingår i $k = 150$-analysen, vilket leder till en större uppskattad variation och bredare kredibilitetsintervall för den mindre analysen. Trots dessa skillnader visade skattningen av $\hat{\mu}$ stabilitet över de olika analysstorlekarna, vilket indikerar robustheten i de valda kombinationerna av a

priori-fördelningar.

Vidare har vi att även om de olika kombinationerna gav liknande punktskattningar för μ, var 95% kredibilitetsintervallen bredast för de likformiga fördelningarna på τ. Detta tyder på att dessa a priori-val för τ kan ha tillåtit en bredare mängd av rimliga värden för effektstorleken. Av denna anledning uteslöts denna fördelning som ett alternativ för τ, och istället valdes halfnormalfördelningen med $\alpha = 0.05$, då den överlag hade de smalaste kredibiltetsintervallen. Ytterligare observerades att valet av fördelning för μ inte hade någon märkbar effekt. Därför beslutade vi att använda normalfördelningen med medelvärde 0 och standardavvikelse 0.5 för μ. De a priori-fördelningar som valdes för den slutliga analysen var följande:

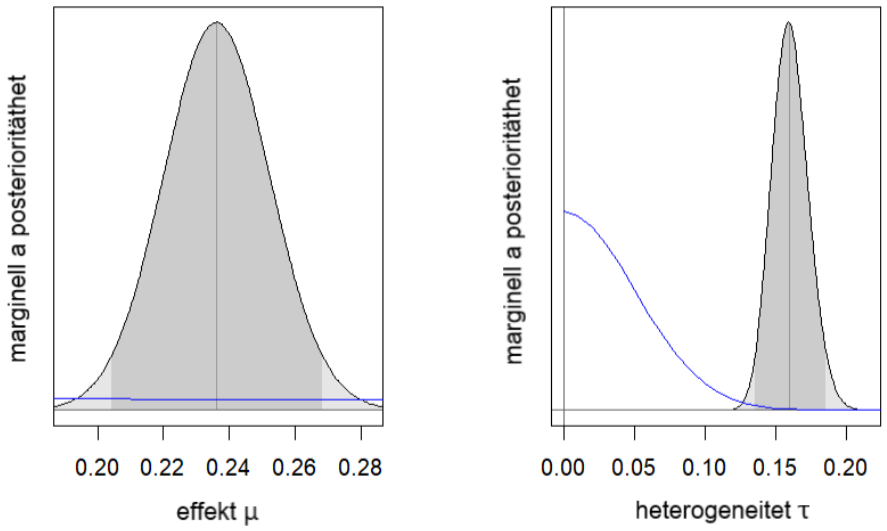
1. $\mu \sim \mathcal{N}(0, 0.5)$
2. $\tau \sim \mathcal{HN}(0.05)$.

En ny metaanalys utfördes med de valda a priorifördelningarna och denna gång inkluderades alla 160 studier. Den resulterande a posteriorimedianen för effektstorleken μ var 0.2362 med ett 95% kredibilitetsintervall på (0.2044, 0.2682). Motsvarande värden för heterogeneiteten τ var 0.1598 med kredibilitetsintervallet (0.1354, 0.1858). Dessa värden omvandlades till korrelationsskalan och vi fick att korrelationen mellan arbetsintervjubedömningar och arbetsprestation var på $\hat{\rho} = 0.2319$. [2]

I figur 5 presenteras de marginella a posterioritätheterna för μ och τ, där ed 95 %-iga kredibilitetsintervallen visas med en mörkare skuggning. A posteriorimedianen markeras med en vertikal linje och i figuren syns även de valda a priorifördelningarna som blå kurvor. I figuren är de två täthetsfunktionerna klockformade. Den horisontella linjen i den nedre delen av grafen representerar a priorifördelningen för μ och indikerar att den var mycket svagt informativ.
I tabell 2 presenteras resultaten från de tre olika metaanalyserna om arbetsintervjuns validitet: McDaniels (1994), frekventistisk och Bayesiansk. Resultaten var konsekventa och gav värdefulla insikter om sambandet mellan arbetsintervjubedömningar och faktiskt arbetsprestation.

[2]Notera: I den frekventistiska metaanalysen användes I^2-statistikan och Cochrans Q-test för att undersöka heterogeniteten bland studierna. Dessa verktyg är dock inte lämpliga för analysen i detta avsnitt, då de bygger på nollhypoteser och p-värden, vilket inte är relevant inom Bayesiansk inferens. istället valdes det att endast rapportera τ som mått på heterogeneitet.

Figur 5: Marginella a posteriori-täthetsfunktioner för μ och τ, tillsammans med deras respektive a priorifördelningar. Skuggningen indikerar 95%-iga kredibilitetsintervall och de blå kurvorna motsvarar a priorifördelningarna.



Tabell 2: Skattningar av korrelations- och heterogeneitetsmått för metaanalyser enligt McDaniel, frekventistisk och Bayesiansk metod.

Mått \ Metaanalys	McDaniel (1994)	Frekventist	Bayesiansk
Korrelation	0.26	0.2325	0.2319
Heterogeneitet	0.17	0.1636	0.1598

4.2.1 Undergruppsanalys

Likt den genomförda analysen i avsnitt 4.1.1 utfördes ytterligare fem metaanalyser för att undersöka korrelationen mellan intervjubedömningar och intervjuform. Dessa analyser utfördes med de tidigare valda a priorivalen ($\mu \sim \mathcal{N}(0, 0.5)$, $\tau \sim \mathcal{HN}(0.05)$), och de skattade värdena av a posteriorimedianerna för ρ och τ^2 presenteras i tabell 3.

De Bayesianska skattningarna för ρ visade sig vara nära de frekventistiska och avvek därmed också från McDaniels resultat. Vidare beräknades heterogeniteten vara lägre i de Bayesianska analyserna, där de strukturerade intervjuerna visade den största skillnaden jämfört med de andra intervjuformerna. För att vidare undersöka värdena av $\hat{\tau}^2$ ritades de marginella a posteriori-täthetskurvorna för heterogeneiteten τ upp för varje intervjuform i figur 6.

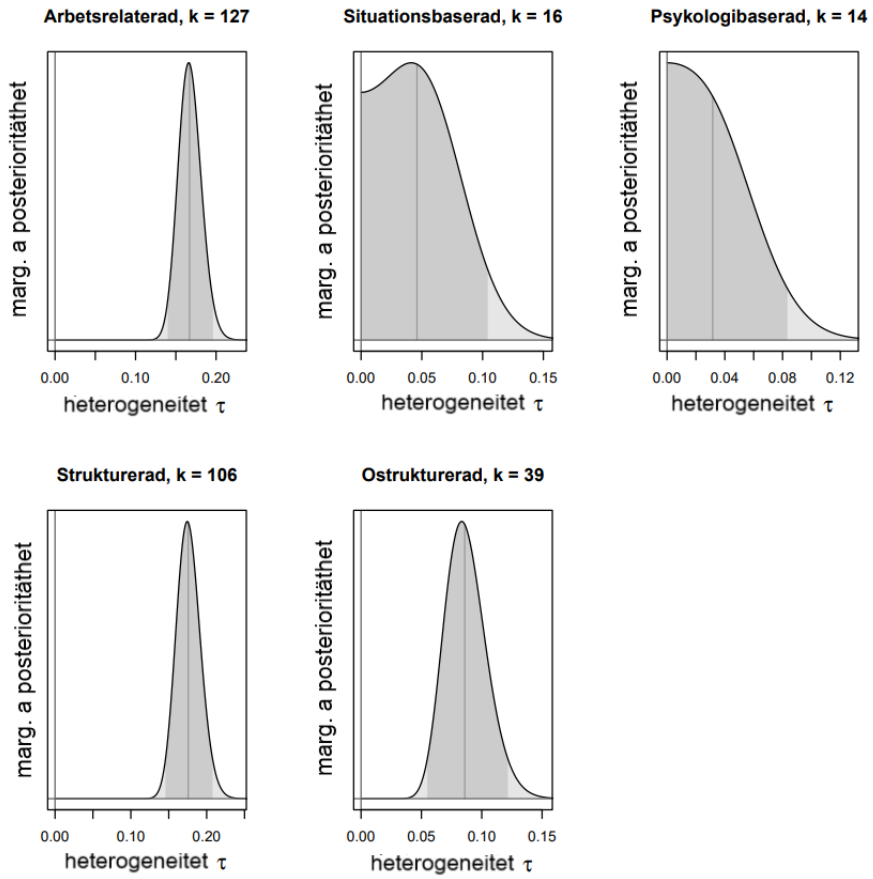
Med ökande antal studier k visade a posteriori-täthetsfunktionerna i figur

Tabell 3: Bayesianska analysresultat av korrelation mellan intervjuer och arbetsprestation, enligt intervjutyp och struktur.

Intervjuform	Antal studier, k	$\hat{\rho}$ (McDaniel)	$\hat{\rho}$	$\hat{\tau}^2$
Arbetsrelaterad	127	0.28	0.2421	0.0274
Situationsbaserad	16	0.35	0.2771	0.0021
Psykologibaserad	14	0.20	0.1500	0.001
Strukturerad	106	0.31	0.2632	0.0302
Ostrukturerad	39	0.23	0.1855	0.0073

6 en tendens att bli smalare och mer klockformade, vilket indikerar en ökad precision i skattningen av heterogeneiteten.

Figur 6: Marginella a posteriori-täthetsfunktioner för τ baserat på intervjuform. Skuggningen indikerar ett 95%-igt kredibilitetsintervall.

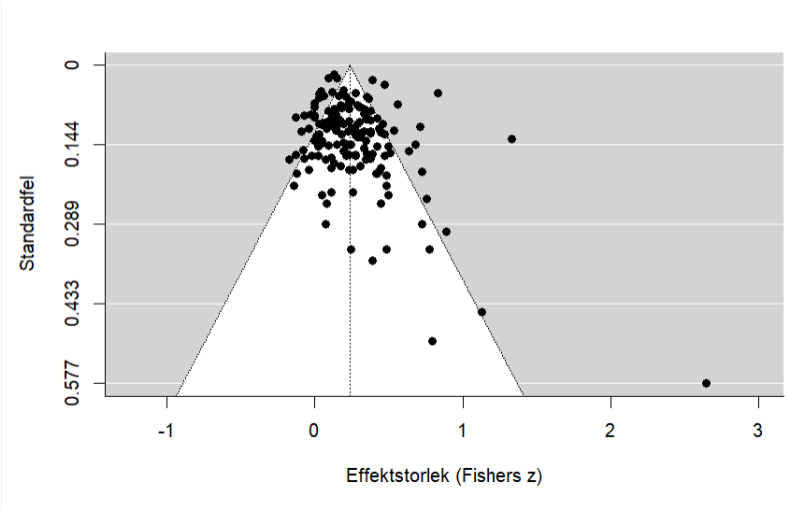


4.3 Publiceringsbias

För att undersöka förekomsten av publiceringsbias bland studierna undersöktes en funnelplott som visade effektstorlekarna i förhållande till deras standardfel, se figur 7. Plotten var centrerad runt medelvärdet av effektstorlekarna (ungefär 0.24) och majoriteten av studierna befann sig i det positiva området. Grafen uppvisade en generell symmetri med en vag skevhet åt höger, och dessutom fanns det få studier med mycket stora standardfel (och därmed små stickprovsstorlekar). Detta antydde att publiceringsbias inte var ett stort problem. Det bör dock noteras att med tanke på att analysen endast innehöll publicerade studier, är det möjligt att det finns några opublicerade studier som kan påverka den skattade effektstorleken.

För att vidare undersöka och korrigera eventuell publiceringsbias genomfördes

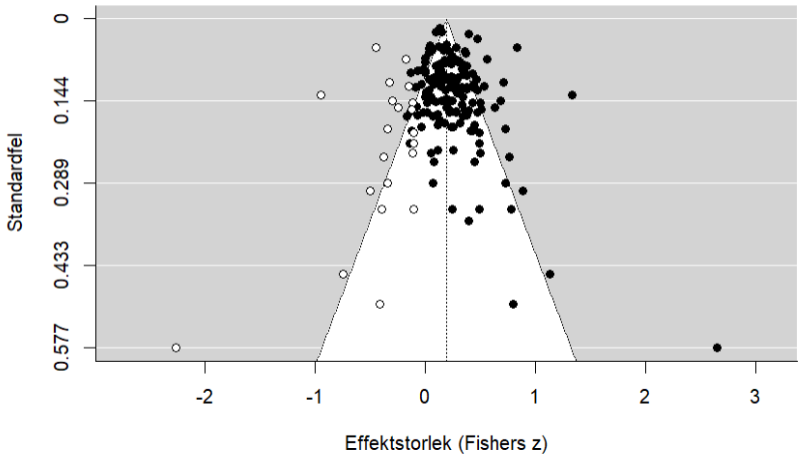
Figur 7: Funnelplott som visar publiceringsbias för de observerade effektstorlekarna. Den triangulära regionen illustrerar ett 95%-igt konfidensintervall.



en trim- och fill-analys. Resultaten av denna analys antydde att det kan ha saknats 21 studier i den vänstra delen av funnelplotten, och efter justering för dessa potentiellt saknade studier genererades en ny funnelplott, se figur 8. Här representeras de nyligen tillagda "studierna" med vita prickar. Den justerade funnelplotten verkade något mer symmetrisk än den ursprungliga, och medelvärdet försköts lite till vänster, vilket indikerar att trim- och fill-analysen kan ha minskat effekten av publikationsbias på våra resultat.

En ny REM metaanalys utfördes där de 21 saknade värdena ("studierna") lades till. Den resulterande skattningen av den genomsnittliga effektstorleken var $\hat{\mu} = 0.1910$ med ett 95% konfidensintervall på (0.1553, 0.2266) i korrelationsskala, och den skattade korrelationen blev $\hat{\rho} = 0.189$. Denna justerade skattning indikerar att jobbintervjuer har en liten till måttlig korrelation med jobbprestation, vilket är i linje med de tidigare resultaten.

Figur 8: Funnelplott som visar publiceringsbias med 95% konfidensintervall efter trim- och fill utförts.



5 Diskussion

I denna studie genomfördes en metaanalys för att undersöka sambandet mellan bedömningar gjorda i anställningsintervjuer och faktisk arbetsprestation. Två olika tillvägagångssätt användes i analysen och resultaten jämfördes med tidigare studier av McDaniel et al. (1994). Resultaten presenteras i tabell 1. Utifrån våra analyser fanns det en skattad korrelation på cirka 0.23. Den frekventistiska analysen resulterade i ett 95% konfidensintervall på (0.2046, 0.2690) för korrelationen. Detta kan tolkas som att om vi utförde analysen oändligt många gånger, skulle intervallet innehålla den sanna korrelationen ρ 95% av fallen. Baserat på detta konfidensintervall och det signifikanta p-värdet för modellen, kan vi **förkasta** nollhypotesen om att det inte finns något samband mellan intervjuer och arbetsprestation med signifikansnivå 95%. Den Bayesianska analysen resulterade i ett 95%-igt kredibilitetsintervall på (0.2044, 0.2682) som i sin tur är mycket likt konfidensintervallet men som tolkas på ett annat sätt. I detta fall är det 95% sannolikhet att den sanna korrelationen ρ ligger inom det givna intervallet. Även detta tyder på att det finns en viss koppling mellan intervjubedömningar och arbetsprestation..

Från våra analyser finner vi att den skattade korrelationskoefficienten är väldigt konsistent mellan de olika metoderna, vilket också gäller för skattningen av τ. Däremot observerade vi en medelhög till hög nivå av heterogeneitet mellan studierna. Skattningen $\hat{\tau} \approx 0.16$ tyder på att variationen i de observerade effektstorlekarna inte enbart kan förklaras av slumpen, utan beror troligen på verkliga skillnader i faktorer som studiedeisgn och studiepopula-

tion. Även I^2-statistikan indikerade en betydande heterogeneitet och Cochrans Q-test var signifikant, vilket ytterligare stöder närvaron av heterogeneitet. En möjlig förklaring till den höga nivån kan vara att bedömningsprocessen för anställningsintervjuer inte är standardiserad mellan olika företag och branscher. Det är möjligt att olika intervjuare tillämpar olika kriterier för att bedöma kandidater, vilket leder till variationer i resultatet av anställningsintervjuer.

De två olika tillvägagångssätten, frekventistisk och Bayesiansk, gav väldigt lika skattningar vilket ger en viss försäkran om att resultaten är robusta och inte alltför beroende av valet av statistisk metod. Däremot skiljer sig metoderna väsentligt, där den frekventistiska är enklare och betydligt mindre beräkningsintensiv. Den Bayesianska metoden kräver en del förarbete med val av a priorifördelningar och deras parametrar. För μ undersöktes två a priorifördelningar som båda betraktas som neutrala och inte beroende på tidigare känd information. Trots deras svaga (eller icke-) informativa karaktär observerades ingen skillnad i skattningarna, som illustreras i figur 4. Detta innebar att båda a priorifördelningarna kunde ses som lämpliga val för att beskriva osäkerheten kring μ.

Valet av a priorifördelning för μ verkade inte spela någon större roll i den slutgiltiga analysen och detta syns tydligt i figur 5. Den marginella a posteriorifördelningen antar en klockformad kurva, medan a priorifördelningen ($\mathcal{N}(0, 0.5)$) för μ representeras av en rak horisontell linje längst ned i figuren. Detta tyder på att datan kan ha starkt påverkat resultatet och att a priorifördelningen har haft liten inverkan, vilket är förväntat vid användning av svaga eller icke-informativa a priorifördelningar. Generellt återspeglar formen på a posteriorifördelningen balansen mellan a priori-kunskap och data (likelihooden). Ett starkt a priori-antagande kan dominera a posteriorifördelningen och leda till slutsatser som inte stöds av datan, medan ett svagt antagande kan domineras av datan och leda till slutsatser som enbart baseras på de observerade värdena.

En möjlig förklaring till a posteriorikurvornas form i figur 5 kan vara storleken på datamängden som användes. Denna analys inkluderade 160 olika studier, vilket kan anses vara en stor mängd data, med tanke på att det oftast räcker med endast ett fåtal (Valentine, Pigott och Hannah R. Rothstein 2010, s. 31). När datamängden är stor (då antalet studier och därmed antalet intervjuer går mot oändligheten), implicerar den centrala gränsvärdessatsen att likelihoodfunktionen vanligtvis konvergerar mot en normalfördelning. Det kan uttryckas som att likelihooden (normalfördelningen) dominerar a priorifördelningen, vilket en klockformad a posteriorifördelning. Samtidigt gäller att när antalet studier (eller datapunkter) ökar, tenderar a posteriorikurvan att bli "smalare" eftersom den ökade mängden data ger mer information. Detta leder till en högre precision i skattningarna. Då a posteriorikurvan representerar sannolikhetsfördelningen för parametern av intresse, givet observerad data och eventuella förkunskaper, blir den mer koncentrerad kring parameterns sanna värde. Detta resulterar i en "tunnare" kurva när mer data samlas in.

Detta framgår tydligt i undergruppsanalysen som genomfördes med Bayesiansk metod. I figur 6 presenteras a posteriorikurvorna för τ för olika intervjuformer, där antalet studier varierar mellan dem. Kurvorna för arbetsrelaterade,

strukturerade och ostrukturerade intervjuer är smala och klockformade, vilket tyder på att tillräckligt många studier ingick i analysen för att ge en noggrann skattning. Å andra sidan skiljer sig kurvorna för situationsbaserade och psykologibaserade intervjuer. Dessa två kurvor är bredare och har formen av en halv-normalfördelning, vilket motsvarar den valda a priorifördelningen för dessa analyserna. Om a posteriorifördelningen är väldigt lik a priorifördelningen, kan det vara en indikation att datan inte har haft ett starkt inflytande på a prioriantagandet. Detta kan uppstå om datan (likelihooden) inte är tillräckligt informativ eller om a prioriantagandet är mycket informativt. I vårt undersökning använde vi en svagt informativ fördelning för τ, vilket tycks ha haft en större påverkan på a posteriorifördelningen. Detta resulterade i att kurvorna för situationsbaserade och psykologibaserade intervjuer fick den karakteristika formen.

Undergruppsanalysen genomfördes med målet att undersöka korrelationen mellan bedömningar gjorda med specifika intervjuformer och faktisk arbetsprestation. Resultaten för de frekventistiska respektive Bayesianska analyserna finns i tabell 1 resp. tabell 3. Resultaten visade att $\hat{\rho}$ antog värden mellan ungefär 0.15 och 0.29 för de olika intervjuformerna, för de två statistiska tillvägagångssätten.

Eftersom intervjuformerna är uppdelade efter både typ och struktur, väljer vi att jämföra skattningarna *inom* de två uppdelningarna istället för *mellan* dem. Detta då en arbetsrelaterad intervju, exempelvis, kan vara antingen strukturerad eller ostrukutrerad. Vi börjar med att undersöka intervjutyp.

De situationbaserade intervjuerna gav den högsta skattningen av den genomsnittliga effektstorleken (Frekventist: $\hat{\rho}_F = 0.2852$, Bayesiansk: $\hat{\rho}_B = 0.2771$) bland de tre intervjutyperna, medan de psykologibaserade hade den lägsta uppskattade effektstorleken ($\hat{\rho}_F = 0.1495$, $\hat{\rho}_B = 0.1500$). Detta tolkas som att situationsbaserade intervjuer är mest effektiva för att förutsäga en kandidats framtida arbetsprestation. Denna slutsats är rimlig då sådana intervjuer kombinerar element från både psykologiska och arbetsrelaterade intervjuer. Situationsbaserade intervjuer syftar till att bedöma hur en kandidat har hanterat specifika situationer eller utmaningar i tidigare arbeten eller erfarenheter. På så sätt bedöms kandidaten både utifrån tidigare arbetslivserfarenheter (som i arbetsrelaterade intervjuer) samt mentala egenskaper och beteende (som i psykologibaserade intervjuer).

När det kommer till heterogeneiteten mellan studier var den som störst bland studier som undersökte arbetsrelaterade intervjuer och lägst bland psykologibaserade. Detta är väntat då de grupperna hade störst respektive lägst antal studier i metaanalysen. En möjlig förklaring till varför τ^2 är störst för arbetsrelaterade intervjuer är att det finns en större variation i den arbetsrelaterade kontexten mellan olika yrken och branscher. Arbetsrelaterade intervjuer fokuserar på kandidaternas kompetenser och erfarenheter som är direkt relevanta för den aktuella tjänsten, och dessa krav och arbetsuppgifter kan skilja sig åt mellan olika yrken och branscher. Detta kan leda till att korrelationen mellan intervjuresultat och arbetsprestation varierar mer mellan olika studier. Å andra sidan är situations- och psykologibaserade intervjuer mer standardiserade och kan användas inom en bredare uppsättning yrken och branscher. I sådana inter-

vjuformer kan intervjuaren exempelvis använda personlighetstester eller andra bedömningsverktyg som är enklare att reproducera och tolka mellan olika yrken, jämfört med att bedöma en kandidat enbart baserat på vad som står på dennes CV. Detta kan resultera i mindre variation i korrelationen mellan intervjuer och arbetsprestation mellan olika studier, vilket i sin tur kan leda till mindre heterogenitet i form av τ^2 i metaanalysen.

Vidare undersöktes påverkan av intervjustruktur på korrelationen. Resultaten visar att det finns en högre korrelation mellan strukturerade intervjuer och arbetsprestation ($\hat{\rho}_F = 0.2659$, $\hat{\rho}_B = 0.2632$) jämfört med ostrukturerade ($\hat{\rho}_F = 0.1862$, $\hat{\rho}_B = 0.1855$). Dessutom är $\hat{\tau}^2$ större för strukturerade intervjuer jämfört med ostrukturerade intervjuer, vilket indikerar en högre heterogenitet i metaanalysen för strukturerade intervjuer.

Skattningarna tyder på att strukturerade intervjuer kan vara en mer pålitlig och effektiv intervjuform för att förutsäga arbetsprestation jämfört med ostrukturerade sådana. Denna observation kan troligtvis förklaras av att strukturerade intervjuer är mer standardiserade och inriktade på specifika arbetsuppgifter och kompetenser, vilket minskar risken för partiskhet och subjektivitet från intervjuarens sida. Dessutom kan ostrukturerade intervjuer vara mer informella och leda till att intervjuare bedömer kandidaterna på ett mer personligt plan, vilket kan påverka reliabiliteten och validiteten hos intervjuerna negativt.

Sammanfattningsvis visade resultaten att bedömningar genom situationsbaserade och strukturerade intervjuer hade den bästa förmågan att förutsäga faktisk arbetsprestation. Det är dock viktigt att notera att skattningarna av ρ och τ^2 för varje intervjuform och intervjustruktur baseras på ett begränsat antal studier, som sträcker sig från 14 till 127 stycken, vilket kan begränsa generaliserbarheten av resultaten. Det är även värt att nämna att våra skattningar skiljer sig från de gjorda av McDaniel. Trots detta kan vi dra slutsatser om vilka intervjuformer som visade lägst till högst korrelation.

Utöver skattningar är det viktigt att diskutera publiceringsbias vid utförandet av en metaanalys. I vår metaanalys har vi endast inkluderat studier som publicerades innan år 1994 och var tillgängliga för nedladdning, vilket kan ha resulterat i exkludering av studier med avvikande resultat. Figur 7 utgör en funnelplot av standardfel mot effektstorlek. En majoritet av studierna samlas kring mittlinjen, som representerar $\hat{\mu}$, och vi observerar en del heterogeneitet bland studierna. Vi ser även en viss asymmetri där fler studier befinner sig till höger om mittlinjen, och att en studie avviker betydligt med ett Fishers z-värde på över 2.5. Den höga effektstorleken i denna studie beror på att den rapporterade ett högt korrelationsvärde, och därmed har den en högre vikt tilldelad i analysen. Vi noterar även att denna studie är väldigt liten (vilket visas med ett högt standardfel) och att den kanske inte har en för stor påverkan på $\hat{\mu}$. Trots att figuren visade en skevhet åt höger, är funnelplotten ganska symmetrisk och majoriteten av studierna faller inom det triangulära 95%-iga konfidensintervallet. Eftersom de flesta studierna ligger inom detta intervall tyder det på låg publiceringsbias.

För att korrigera resultatet utfördes en trim- och fill-analys som lade till 21 saknade (hypotetiska) studier. Med hjälp av dessa beräknades ett nytt värde på $\hat{\mu}$ och vi fick en korrelation på $\hat{\rho} \approx 0.19$. Denna skattning avvek något från

den tidigare skattningen, men inte i hög grad. Det är viktigt att notera att även om korrigeringen av funnelplotten minskade effekten av publiceringsbias, kan det fortfarande finnas faktorer som påverkar generaliserbarheten och tillförlitligheten hos resultaten. Sådana faktorer kan inkludera selektiv rapportering av resultat och brister i kvaliteten hos de studier som inkluderades i analysen.

Vi har nu skattat korrelationen på tre olika sätt: frekventistiskt, Bayesianskt och frekventistiskt med korrigering för publiceringsbias. De två förstnämnda skattningarna ($\hat{\rho} \approx 0.23$) liknar den som McDaniel rapporterade ($\hat{\rho} = 0.26$), medan den tredje inte kan jämföras direkt eftersom McDaniel inte tog hänsyn till publiceringsbias i sin analys. En möjlig förklaring till avvikelsen mellan våra skattningar och McDaniels kan vara att han använde Hunter-Schmidts metod (Hunter och Schmidt 2004), medan vår omvärdering använder formler och procedurer som beskrivs av Hedges och Olkin (Hedges och Olkin 1985). Hedges-Olkins metod bygger på en modell där variationen mellan effektstorlekarna orsakas av både verklig heterogenitet mellan studierna och slumpmässig variation. Denna metod använder en viktning som baseras på varje studies precision, samtidigt som den är känd för att kunna hantera små stickprov och data som kan vara missvisande på grund av publiceringsbias. I jämförelse använder Hunter-Schmidts metod en annan formel för att beräkna den gemensamma effektstorleken och en annan viktning som bygger på varje studies storlek. Detta tillvägagångssätt är mer lämpligt för att hantera heterogenitet i effektstorlekarna mellan studierna och är mindre känsligt för missvisande data. Överlag föredras Hedges-Olkins metod för REM (Johnson, Mullen och Salas 1995, s. 12). Att olika formler använts för beräkningen kan vara orsaken till att skattningen skiljer sig med en decimal. I övrigt var skattningarna ganska nära varandra.

Även om den genomförda metaanalysen i denna uppsats ger värdefulla insikter om sambandet mellan variablerna arbetsintervjuer och arbetsprestation, finns det flera sätt att förbättra och utvidga analysen. Det är särskilt värt att reflektera över de studier som inkluderats i analysen, då samtliga är från innan 1994. En möjlig förbättring och utökning av analysen hade varit att ha med studier gjorda under senare tid för att undersöka om korrelationen har förändrats över tid.

6 Slutsats

Baserat på resultaten av vår metaanalys visar dessa att det finns en signifikant korrelation mellan bedömningar gjorda under anställningsintervjuer och faktisk arbetsprestation. Vi har undersökt detta samband genom att tillämpa två olika analytiska tillvägagångssätt och har skattat korrelationskoefficienten till cirka 0.23 med ett 95% konfidensintervall på (0.2018, 0.2627) för den frekventistiska metoden, och ett 95% kredibilitetsintervall på (0.2020, 0.2620) för den Bayesianska metoden. Våra resultat överensstämmer till stor del med McDaniels analys från 1994, men det förekommer vissa mindre skillnader, vilket kan förklaras med att olika statistiska metoder användes.

Den betydande heterogeniteten mellan studier tyder på att det finns en bristande standardisering i processen att utvärdera jobbkandidater i intervjuer, vilket sträcker sig över olika företag och branscher. Både de frekventistiska och Bayesianska analyserna gav liknande resultat, vilket stärker förtroendet för robustheten av våra resultat och att valet av statistisk metod inte påverkar den övergripande slutsatsen.

Sammanfattningsvis visar vår metaanalys att anställningsintervjubedömningar kan ge värdefull information om arbetsprestation. Men på grund av heterogeniteten mellan studierna rekommenderar vi ytterligare forskning för att identifiera bästa praxis för utvärdering av jobbkandidater under intervjuer.

Referenser

Bishop, Christopher M och Nasser M Nasrabadi (2006). *Pattern recognition and machine learning*. Vol. 4. 4. Springer.

Borenstein, Michael m. fl. (2009). *Introduction to meta-analysis*. John Wiley & Sons.

— (2010). "A basic introduction to fixed-effect and random-effects models for meta-analysis". I: *Research synthesis methods* 1.2, s. 97–111.

Carlin, Bradley P. och Thomas A. Louis (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. 2nd. Chapman och Hall/CRC.

Casella, G och RL Berger (2002). *Statistical Inference*. 2nd. Cengage Learning.

Cover, Thomas M och Joy A Thomas (1991). *Information theory and statistics*. 2nd. John Wiley Sons, ltd.

Dettori, Joseph R., Daniel C. Norvell och Jens R. Chapman (2022). "Fixed-Effect vs Random-Effects Models for Meta-Analysis: 3 Points to Consider". I: *Global Spine Journal* 12.7. PMID: 35723546, s. 1624–1626. DOI: 10.1177/21925682221110527. URL: https://doi.org/10.1177/2192568222111052.

Egger, Matthias m. fl. (1997). "Bias in meta-analysis detected by a simple, graphical test". I: *Bmj* 315.7109, s. 629–634.

Hedges, Larry V och Ingram Olkin (1985). *Statistical methods for meta-analysis*. Academic press.

Held, Leonhard och Daniel Sabanés Bové (2014). *Applied statistical inference*. Vol. 10. 978-3. Springer.

Hesser, Hugo och Gerhard Andersson (2015). *Introduktion till metaanalys och systematiska översikter*. Studentlitteratur AB.

Higgins, Julian PT och Simon G Thompson (2002). "Quantifying heterogeneity in a meta-analysis". I: *Statistics in medicine* 21.11, s. 1539–1558.

Higgins, Julian PT, Simon G Thompson m. fl. (2003). "Measuring inconsistency in meta-analyses". I: *Bmj* 327.7414, s. 557–560.

Hunter, John E och Frank L Schmidt (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage.

Johnson, Blair T, Brian Mullen och Eduardo Salas (1995). "Comparison of three major meta-analytic approaches." I: *Journal of applied psychology* 80.1, s. 94.

Klein, Nadja och Thomas Kneib (2016). "Scale-dependent priors for variance parameters in structured additive distributional regression". I.

Kowalski, Charles J. (1972). "On the Effects of Non-Normality on the Distribution of the Sample Product-Moment Correlation Coefficient". I: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 21.1, s. 1–12. URL: `http://www.jstor.org/stable/2346598` (hämtad 2023-04-13).

McDaniel, Michael A m. fl. (1994). "The validity of employment interviews: A comprehensive review and meta-analysis." I: *Journal of applied psychology* 79.4, s. 599.

Olson, Carin M m. fl. (2002). "Publication bias in editorial decision making". I: *Jama* 287.21, s. 2825–2828.

Rivera, Lauren A. (2012). "Hiring as Cultural Matching: The Case of Elite Professional Service Firms". I: *American Sociological Review* 77.6, s. 999–1022. DOI: `10.1177/0003122412463213`. eprint: `https://doi.org/10.1177/0003122412463213`. URL: `https://doi.org/10.1177/0003122412463213`.

Rosenthal, Robert (1979). "The file drawer problem and tolerance for null results." I: *Psychological bulletin* 86.3, s. 638.

Rothstein, Hannah R, Alexander J Sutton och Michael Borenstein (2005). "Publication bias in meta-analysis". I: *Publication bias in meta-analysis: Prevention, assessment and adjustments*, s. 1–7.

Röver, Christian (2020). "Bayesian Random-Effects Meta-Analysis Using the bayesmeta R Package". I: *Journal of Statistical Software* 93.6, s. 1–51. DOI: `10.18637/jss.v093.i06`. URL: `https://www.jstatsoft.org/index.php/jss/article/view/v093i06`.

Röver, Christian och Tim Friede (2017). "Discrete approximation of a mixture distribution via restricted divergence". I: *Journal of Computational and Graphical Statistics* 26.1, s. 217–222.

Schulz, Susanne m. fl. (2021). "Systematically Defined Informative Priors in Bayesian Estimation: An Empirical Application on the Transmission of Internalizing Symptoms Through Mother-Adolescent Interaction Behavior". I: *Frontiers in Psychology* 12. DOI: `10.3389/fpsyg.2021.620802`. URL: `https://www.frontiersin.org/articles/10.3389/fpsyg.2021.620802`.

Sears, Greg J och Patricia M Rowe (2003). "A personality-based similar-to-me effect in the employment interview: Conscientiousness, affect-versus competence-mediated interpretations, and the role of job relevance." I: *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement* 35.1, s. 13.

Seidel, Wilfried (2011). "Mixture Models". I: *International Encyclopedia of Statistical Science*. Utg. av Miodrag Lovric. Berlin, Heidelberg: Springer Berlin Heidelberg, s. 827–829. ISBN: 978-3-642-04898-2. DOI: 10.1007/978-3-642-04898-2_368. URL: https://doi.org/10.1007/978-3-642-04898-2_368.

Sterne, Jonathan A.C och Matthias Egger (2001). "Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis". I: *Journal of Clinical Epidemiology* 54.10, s. 1046–1055. ISSN: 0895-4356. DOI: https://doi.org/10.1016/S0895-4356(01)00377-8. URL: https://www.sciencedirect.com/science/article/pii/S0895435601003778.

Thornton, Alison och Peter Lee (2000). "Publication bias in meta-analysis: its causes and consequences". I: *Journal of clinical epidemiology* 53.2, s. 207–216.

Tufanaru, Catalin m. fl. (2015). "Fixed or random effects meta-analysis? Common methodological issues in systematic reviews of effectiveness". I: *JBI Evidence Implementation* 13.3, s. 196–207.

Valentine, Jeffrey C., Therese D. Pigott och Hannah R. Rothstein (2010). "How Many Studies Do You Need?: A Primer on Statistical Power for Meta-Analysis". I: *Journal of Educational and Behavioral Statistics* 35.2, s. 215–247. DOI: 10.3102/1076998609346961. URL: https://doi.org/10.3102/1076998609346961.

Williams, Donald R, Philippe Rast och Paul - Christian Bürkner (jan. 2018). *Bayesian Meta-Analysis with Weakly Informative Prior Distributions*. DOI: 10.31234/osf.io/7tbrm. URL: psyarxiv.com/7tbrm.