Metodologi for replikasjonsprosjekt

Ole Fredrik Borgundvåg Berg
NTNU

Dette er et replikasjonsprosjekt som er basert på et tilsvarende prosjekt som ble gjort av Ulrich Schimmack, der han prøvde å finne i hvilken grad vitenskapelige funn ved ulike amerikanske universiteter kunne replikeres, inkludert estimert replikasjonsrate på enkeltforskere (Schimmack, 2022). Siden faktisk replikasjon er dyrt, så laget heller han med flere en modell som finner estimert replikasjonsrate basert på p-verdier i artiklene til forskerne. Modellen er implementert i biblioteket «z-curve» tilgjengelig for programvaren R (Bartoš & Schimmack, 2022; Brunner & Schimmack, 2020). I dette prosjektet har jeg brukt samme modell for å finne estimert replikasjonsrate på forskere ved Institutt for Psykologi ved NTNU.

Innhenting av artikler

For å finne artikler og forskere ble Cristin APIet brukt ("Cristin API | Cristin REST API", udatert). APIet gjør det enkelt både å finne ut hvilke forskere som er tilknyttet hvilke institutt og hvilke artikler som er skrevet av hvilke forskere. Artiklene som er brukt i dette prosjektet er de publikasjonene som har `category.code = "ARTICLE"` i APIet. Nedlastingen av artiklene ble også gjort programmerisk, men ikke ved hjelp av Cristin. Dette kan føre til at noen av artiklene i Cristin ikke kom med i dette prosjektet pga. manglende tilgang.

Ekstrahering av z-verdier

Siden p-verdier ofte er rapportert på formen $p < 0.05$ er det vanskelig å lage et histogram ut fra disse. Derfor bygget analysen kun på data fra statistiske tester hvor mer eksakte test-statistikker ble rapportert. Deretter konverteres test-statistikken til en z-verdi med tilsvarende p-verdi. Selve ekstraheringen av de statistiske testene ble gjort ved hjelp av verktøyet JATSdecoder og funksjonen `get.stats()` (Böschen, 2021). Z-curve har også funksjoner for å ekstrahere statistiske tester, men ifølge egen testing og Böschen (2021) sine forsøk gir JATSdecoder bedre resultater i ekstraheringen av teststatistikker, sammenlignet med z-curve.

Modellering av replikasjonsrate og visualisering

For å finne forventet replikasjonsrate og for å lage grafer for visualisering, ble z-verdiene puttet inn i z-curve. z-curve både regnet ut forventet replikasjonsrate og laget histogram med z-verdiene. For mer informasjon om hvordan z-curve fungerer, se Brunner og Schimmack (2020) og Bartoš og Schimmack (2022).

Kode

Koden som ble brukt i dette prosjektet er tilgjengelig på https://github.com/OleFredrik1/Replikasjonsprosjekt.

Hva betyr de ulike begrepene?

En hypotese er en påstand om noe og er det man skal teste empirisk. Man opererer med en nullhypotese og en alternativ hypotese. Nullhypotesen er den mest konservative og som gjør færrest mulige antagelser. Typisk er nullhypotesen at det ikke er noen sammenheng mellom to eller flere variabler, mens den alternative hypotesen sier at det er en sammenheng.

En p-verdi er sannsynligheten for at man observerer noe som er like eller mer ekstremt enn de dataene man faktisk får, gitt at nullhypotesen er sann. Merk at dette er ikke det samme som sannsynligheten for at nullhypotesen er sann.

Om p-verdien man får er lavere enn et forhåndsbestemt signifikansnivå sier man at resultatet er statistisk signifikant og man forkaster nullhypotesen. I psykologi (og en del andre fagfelt) er det standard å bruke 0.05 som signifikansnivå. Merk at dersom nullhypotesen er sann er det fortsatt 5% sjanse for at man får et signifikant resultat og forkaster nullhypotesen.

En z-verdi er en normalfordelt verdi. Vi har at $\Pr[|Z| \leq 1.96] \approx 0.05$ når $Z \sim N(0, 1)$, eller med andre ord vil en z-verdi med absoluttverdi på 1.96 eller mer vil ha en p-verdi på 0.05 eller mindre.

Hva er p-hacking?

P-hacking når man gjør en rekke triks for å prøve å få lavere p-verdier. Det finnes mange ulike måter å p-hacke på. En måte er å prøve ulike statistiske tester og så velge de som gir lavest p-verdier. Siden ulike statistiske tester varierer litt i hvordan de funker kan de gi litt ulike p-verdier. En annen måte man kan p-hacke er å uten god grunn fjerne observasjoner som ikke passer til hypotesen ved f.eks. å si at de er uteliggere. En tredje måte er å endre modellen man bruker helt til man får signifikante resultater. Ved å endre nok kan man ende opp med en signifikant p-verdi til slutt selv om nullhypotesen er sann. Noe annet kan være å se på subgrupper dersom ikke effekten stemmer for hele populasjonen. Kanskje det stemmer bare for kvinner, de mellom 40-50 år eller de med høy inntekt. Hvis du leter nok, kan du nok finne en subgruppe som gir deg lav nok p-verdi. Alt dette er forskningsjuks dersom man ikke justerer for multippel hypotesetesting, men heller bare publiserer de signifikante resultatene og forkaster resten uten å nevne det.

En interessant anekdote er den studien som fant sammenheng mellom fullmåne og selvmord, men bare hos kvinner under 45 år og bare på vinteren (Meyer-Rochow mfl., 2021). Denne klarte merkelig nok ikke å replikeres (Plöderl mfl., 2023).

Hva er publiseringsbias?

Om det kun er 5% sjanse for at man forkaster en sann nullhypotese, betyr det da at om man søker etter artikler som tester en usann hypotese så vil 95% av artiklene ha nullresultater? Ikke nødvendigvis. Det er en tendens til at artikler som ikke har funnet noe signifikant blir publisert sjeldnere enn artikler som har funnet noe signifikant. Dette kalles

publiseringsbias. I verste tilfelle, hvis ingen av nullresultatene publiseres, men kun de 5% som får signifikante resultater, så vil alle artiklene på område si at hypotesen stemmer selv om den ikke gjør det.

Grunnene til publiseringsbias er flerfoldige. Noe skyldes at det gir mer prestisje å publisere signifikante resultater enn nullresultater og om man får nullresultater kan det være fristende å heller legge resultatene i en skuff og starte et nytt forsøk, heller enn å måtte skrive artikkel og publisere dem. Det kan også være at journaler har mer lyst til å publisere signifikante resultater enn nullresultater, fordi de signifikante resultatene anses å være mer interessante. Det kan også være at forskere ikke publiserer nullresultater fordi man tror man har gjort noe feil, spesielt hvis nullresultatet skiller seg ut fra forskningen på området ellers.

Når man gjør meta-analyser må man justere for publiseringsbias. Om man ikke gjør det går det veldig feil. En studie fant at i psykologi ville man sett at en median sannsynlighet på 98.9% for at en meta-analyse ville støtte hypotesen om man ikke justerte for publiseringsbias, mens om man justerte for publiseringsbias ville sannsynligheten reduseres til 55.7% (Bartoš mfl., 2022). Altså fra å nesten helt sikkert støtte hypotesen, til å ende opp med kron og mynt.

En annen effekt er at forskning som ikke kan replikeres siteres oftere, muligens fordi de har mer spennende funn som dessverre viser seg å være for gode til å være sanne (Serra-Garcia & Gneezy, 2021).

Hvorfor det meste av forskning ikke stemmer

Ioannidis (2005) sin artikkel har det litt alarmerende navnet «Why Most Published Research Findings Are False». En faktor som ikke er nevnt til nå er hva som er apriori sannsynlighet for at en hypotese stemmer? Se for deg at du har et lykkehjul med variabler, som for kjønn, alder, inntekt, utdanning, introversjon, reisevaner, klesvalg, bruk av et bestemt legemiddel, depressive symptomer, vel alt du kan tenke deg. Deretter spinner man dette hjulet to ganger for å finne to variabler A og B. Hva er da sannsynligheten for at det er noen sammenheng mellom de to variablene? Ioannidis (2005) argumenterer med at det i mange tilfeller er ganske lav sjanse. Om T er hendelsen at hypotesen stemmer, F er hendelsen at hypotesen ikke stemmer og $+$ er hendelsen at man får et signifikant resultat, så kan vi finne $\Pr[T|+]$, altså sannsynligheten for at en hypotese stemmer gitt et signifikant resultat, ved å bruke Bayes' teorem:

$$\Pr[T|+] = \frac{\Pr[+|T]\Pr[T]}{\Pr[+]}$$

Hvor $\Pr[+|T]$ er sannsynligheten for at man får et signifikant resultat gitt at hypotesen stemmer og kalles ofte styrken til testen. Man setter typisk opp forsøket slik at denne verdien er antatt å være 0.8. $\Pr[+]$ kan utvides til $\Pr[+|T]\Pr[T] + \Pr[+|F]\Pr[F]$, hvor $\Pr[+|F]$ er signifikansnivået vårt 0.05. Om man setter $\Pr[T]$ til å være 0.1 (altså én av ti hypoteser stemmer), vil man få en sannsynlighet på $\frac{0.8\times0.1}{0.8\times0.1+0.05\times0.9} = 0.64$ for at en hypotese stemmer gitt at man får et signifikant resultat. Dette vil tilsi at rundt to av tre signifikante resultater vil tilsi at hypotesen stemmer. Hvis man legger til publiseringsbias i likningen blir sannsynligheten enda lavere. Ioannidis (2005) argumenterer for at denne sannsynligheten er under 50% i mange tilfeller, derav tittelen på artikkelen.

Rigtig nok har Ioannidis (2005) har fått en del kritikk av andre forskere, blant annet skaperen av z-curve, som mener at Ioannidis sin påstand om at det meste av forskning er feil ikke stemmer overens med dataene man har (Schimmack, 2019).

Referanser

Bartoš, F., Maier, M., Wagenmakers, E.-J., Nippold, F., Doucouliagos, H., Ioannidis, J. P. A., Otte, W. M., Sladekova, M., Deresssa, T. K., Bruns, S. B., Fanelli, D., & Stanley, T. D. (2022). Footprint of Publication Selection Bias on Meta-Analyses in Medicine, Environmental Sciences, Psychology, and Economics. https://doi.org/10.48550/arXiv.2208.12334

Bartoš, F., & Schimmack, U. (2022). Z-Curve 2.0: Estimating Replication Rates and Discovery Rates. *Meta-Psychology*, *6*. https://doi.org/10.15626/MP.2021.2720

Brunner, J., & Schimmack, U. (2020). Estimating Population Mean Power Under Conditions of Heterogeneity and Selection for Significance. *Meta-Psychology*, *4*. https://doi.org/10.15626/MP.2018.874

Böschen, I. (2021). Evaluation of JATSdecoder as an automated text extraction tool for statistical results in scientific reports. *Scientific Reports*, *11*(1), 19525. https://doi.org/10.1038/s41598-021-98782-3

Cristin API | Cristin REST API. (udatert). Hentet 22. januar 2023, fra https://api.cristin.no/

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Meyer-Rochow, V. B., Hakko, T., Hakko, H., Riipinen, P., & Timonen, M. (2021). Synodic lunar phases and suicide: Based on 2605 suicides over 23 years, a full moon peak is apparent in premenopausal women from northern Finland. *Molecular Psychiatry*, *26*(9), 5071–5078. https://doi.org/10.1038/s41380-020-0768-7

Plöderl, M., Westerlund, J., Hökby, S., Hadlaczky, G., & Hengartner, M. P. (2023). Increased suicide risk among younger women in winter during full moon in northern Europe. An artifact or a novel finding? *Molecular Psychiatry*, *28*(2), 901–907. https://doi.org/10.1038/s41380-022-01823-0

Schimmack, U. (2019). Ioannidis (2005) was wrong: Most published research findings are not false. Hentet 23. januar 2023, fra https://replicationindex.com/2019/01/15/ioannidis-2005-was-wrong-most-published-research-findings-are-not-false/

Schimmack, U. (2022). Replicability Rankings of Psychology Departments. Hentet 8. januar 2023, fra https://replicationindex.com/2022/03/15/rr22-psy-dept/

Serra-Garcia, M., & Gneezy, U. (2021). Nonreplicable Publications Are Cited More than Replicable Ones. *Science Advances*, *7*(21), eabd1705. https://doi.org/10.1126/sciadv.abd1705