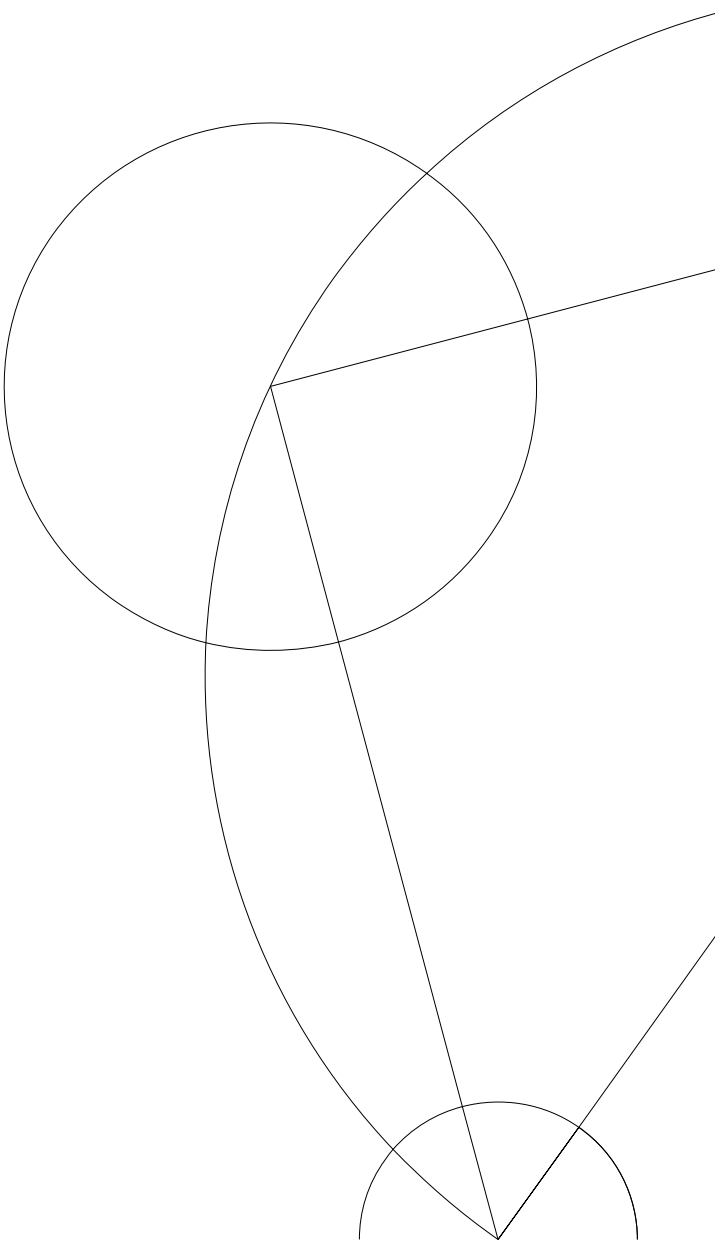
KØBENHAVNS UNIVERSITET
INSTITUT FOR NATURFAGENES DIDAKTIK

P-hacking

Mille Bødstrup
Bachelor i matematik



Vejleder: Mikkel Willum Johansen

22. Marts 2024

IND's studenterserie nr. 118, 2024

INSTITUT FOR NATURFAGENES DIDAKTIK, www.ind.ku.dk
Alle publikationer fra IND er tilgængelige via hjemmesiden.

IND's studenterserie

87. Jesper Jul Jensen: Formativ evaluering og faglige samspil i almen studieforberedelse (2020)
88. Karen A. Voigt: Assessing Student Conceptions with Network Theory - Investigating Student Conceptions in the Force Concept Inventory Using MAMCR (2020)
89. Julie Hougaard Overgaard: Using virtual experiments as a preparation for large scale facility experiments (2020)
90. Maria Anagnostou: Trigonometry in upper secondary school context: identities and functions (2020)
91. Henry James Evans: How Do Different Framings Of Climate Change Affect Pro-environmental Behaviour? (2020)
92. Mette Jensen: Study and Research Paths in Discrete Mathematics (2020)
93. Jesper Hansen: Effekten og brugen af narrative læringsspil og simuleringer i gymnasiet (2020)
94. Mie Haumann Petersen: Bilingual student performance in the context of probability and statistics teaching in Danish High schools (2020)
95. Caroline Woergaard Gram: "Super Yeast" - The motivational potential of an inquiry-based experimental exercise (2021)
96. Lone Brun Jakobsen: Kan man hjælpe elevers forståelse af naturvidenskab ved at lade dem formulere sig om et naturvidenskabeligt emne i et andet fag? (2021)
97. Maibritt Oksen og Morten Kjøller Hegelund: Styrkelse af motivation gennem Webinar og Green Screen (2021)
98. Søren Bystrup Jacobsen: Peer feedback: Fra modstand til mestring? (2021)
99. Bente Guldbrandsen: Er der nogen, som har spurgt en fysiklærer? (2021)
100. Iben Vernegren Christensen: Bingoplader i kemiundervisningen – en metode til styrkelse af den faglige samtale? (2021)
101. Claus Axel Frimann Kristinson Bang: Probability, Combinatorics, and Lesson Study in Danish High School (2021)
102. Derya Diana Cosan: A Diagnostic Test for Danish Middle School Arithmetics (2021)
103. Kasper Rytter Falster Dethlefsen: Formativt potentiale og udbytte i Structured Assessment Dialogue (2021)
104. Nicole Jonasen: A diagnostic study on functions (2021)
105. Trine Nørgaard Christensen: Organisatorisk læring på teknisk eux (2021)
106. Simon Funch: Åben Skole som indgang til tværfagligt samarbejde (2022)
107. Hans-Christian Borggreen Keller: Stem som interdisciplinær undervisningsform (2022)
108. Marie-Louise Krarup, Jakob Holm Jakobsen, Michelle Kyk & Malene Hermann Jensen: Implementering af STEM i grundskolen (2022)
109. Anja Rousing Lauridsen & Jonas Traczyk Jensen: Grundskoleelevers oplevelse af SSI-undervisning i en STEM-kontekst. (2022)
110. Aurora Olden Aglen: Danish upper secondary students' apprehensions of the equal sign (2023)
111. Metine Rahbek Tarp & Nicolaj Pape Frantzen: Machine Learning i gymnasiet (2023)
112. Jonas Uglebjerg: Independence in Secondary Probability and Statistics: Content Analysis and Task Design (2023)
113. Hans Lindebjerg Legard: Stopmotion som redskab for konceptuel læring. (2023)
114. Caroline Woergaard Gram & Dan Johan Kristensen: The ice algae Ancylonema as icebreakers: A case study on how the international Deep Purple Research Project can create meaningful outreach in Greenland. (2023)
115. Julie Sloth Bjerrum: 'KLIMA HISTORIER' The Art Of Imagining A Green Future. (2023)
116. Emilie Skaarup Bruhn: Muligheder og udfordringer ved STEM-undervisning (2024)
117. Milla Mandrup Fogt: Undersøgelsesbaseret undervisning i Pascals trekant (2024)
118. **Mille Bødstrup: P-hacking (2024)**

IND's studenterserie *omfatter kandidatspecialer, bachelorprojekter og masterafhandlinger skrevet ved eller i tilknytning til Institut for Naturfagenes Didaktik. Disse drejer sig ofte om uddannelsesfaglige problemstillinger, der har interesse også uden for universitetets mure. De publiceres derfor i elektronisk form, naturligvis under forudsætning af samtykke fra forfatterne. Det er tale om studenterarbejder, og ikke endelige forskningspublikationer.*
Se hele serien på: www.ind.ku.dk/publikationer/studenterserien/

Abstract

This bachelor project explores the phenomenon of p-hacking within scientific research - a term describing the practice where researchers manipulate data to achieve statistically significant results. Through an analysis of three case studies (including Brian Wansink's pizza buffet study, an investigation of motorcycle accidents during full moons and Study 329 on the use of Paroxetine in adolescents) it clarify the underlying causes, methods employed, and the potentially consequences of p-hacking. By utilizing case studies as the methodological approach, the thesis aims to uncover the often subtle and overlooked choices contributing to p-hacking, as well as discussing its impact on the integrity and credibility of science. The project includes a mathematical walkthrough of the theory behind hypothesis testing and p-values, to give a better understanding of the issue. Finally, potential methods to prevent p-hacking are discussed - emphasizing the importance of transparency and accountability in the research process.

Indhold

1 Indledning

Vi lever i et samfund, hvor man bliver konfronteret med virkelig mange informationer hver eneste dag. I TV'et står vejrværter i gummistøvler og rapporterer om stormen - inde i stormen, mens smartphonen konstant lyser op af notifikationer om primærvalget i USA og konflikten i Gaza. I dette massive hav af informationer, støder man ofte på artikler baseret på et videnskabeligt studie, som konkluderer en eller anden korrelation mellem en bestemt adfærd og en tilhørende konsekvens. Et hurtigt klik ind på www.videnskab.dk, og man har inden for de første 20 sekunder fundet en artikel, der oplyser, at "natteravne har næsten dobbelt så stor risiko for åreforkalkning som morgenmennesker" (Jakobsen, 2024). Rubrikken er baseret på et studie fra Göteborg Universitet, og er et godt eksempel på et forsningsresultat, som er blevet publiceret, videreformidlet og herefter læst af en almen dansker.

Kigger man nærmere på studiet fra Göteborg, får man et indblik i forskernes arbejdsproces. De beskriver, hvordan data er blevet indsamlet, hvilket data de har valgt - eller udeladt, og hvordan de herefter er blevet testet og fortolket. Man kan se, at der tages udgangspunkt i flere forskellige variable - både kategoriske variable som bla. køn, uddannelsesniveau og "graden af b-menneske", men også numeriske variable som BMI, alder og kolesteroltal. De forklarer desuden, at de udregnede p-værdier er beregnet ud fra en χ^2-test og en t-test for hhv. for kategoriske og numeriske variable, i øvrigt med et signifikationsniveau på 0.05. (Bergström et al., 2024)

Arbejdsmetoden i ovenstående studie, som i øvrigt er gældende for langt de fleste studier, involverer en formulering af en nulhypotese og en alternativ hypotese. Herefter udregnes en p-værdi ud fra en bestemt teststørrelse for netop at vurdere styrken af evidensen mod nulhypotesen. P-værdien bliver brugt flittigt, idet den er afgørende for at bestemme, om forskningsresultaterne kan betragtes som statistisk signifikante. P-værdien spiller derfor en central rolle i fortolkningen af data - og det er netop denne p-værdi, som vil være omdrejningspunktet i dette projekt. Spørgsmålet er nemlig, om man gøre noget for at manipulere med denne værdi?

Svaret er ja, og det kaldes *p-hacking*.

Tankerne leder en hen på en hætteklædt person, der sidder i et mørkt lokale bag en computerskærm - men det er altså ikke den form for hacking. P-hacking er nemlig en meget subtil størrelse, som opstår, når forskere utilsigtet eller bevidst manipulerer med p-værdien for at opnå et statistisk signifikant resultat - og det manifesterer sig i de valg, som forskerne træffer undervejs i forskningsprocessen.

I studiet fra Göteborg skulle forskerne også træffe mange valg. Det kunne være spørgsmål, som: Hvordan udvælges forsøgspersonerne - og hvor mange? Hvilke personer inkluderes eller ekskluderes? Hvordan defineres graden af b-menneske? Hvilke parametre skal der måles på? Hvilken teststørrelse skal benyttes? Listen af valg kan være *meget* lang - og disse valg kan potentielt medføre p-hacking.

Vi kunne nemlig forestille os, at Bergström et al. (2024) havde valgt løbende at indsamle data, og først stoppede, når p-værdien lige akkurat kom under 0.05 - eller at der var en overrepræsentation af rygere og ældre personer i deres datasæt, fordi de havde valgt at udelukke dem, der ikke var det. Det kunne også tænkes, at deres oprindelige hypotese var, at b-mennesker havde højere BMI eller kolesteroltal, eller at a-mennesker havde større tendens til at ryge end b-mennesker - men

at disse hypoteser ikke havde vist noget signifikant resultat. Derefter kunne de have valgt kun at rapportere deres signifikante resultat, og måske bevidst eller utilsigtet have udeladt at fortælle om alle de variabler eller hypoteser, som ikke gav et resultat.

Hvis Bergström et al. (2024) havde truffet disse valg, ville konklusionen om, at b-mennesker har større tendens til åreforkalkning, nok være misvisende - for disse valg ville betyde, at der var blevet manipuleret med p-værdien, hvorfor det kan betragtes som p-hacking.

Dette projekt har til formål at undersøge de forskellige aspekter inden for p-hacking, herunder:

- Hvordan opstår p-hacking (hvilke metoder)?
- Hvorfor opstår det (hvilke motiver)?
- Hvilke konsekvenser har det - og hvordan kan vi mindske det?

Disse spørgsmål vil blive undersøgt og forhåbentlig besvaret med hjælp fra tre forskellige cases, som alle repræsenterer noget vidt forskelligt. Dette vil lede op til en diskussion om videnskabelig integritet - om ansvar og etik, samt hvilke metoder, vi kan tage i brug for at forhindre det.

Projektet skulle derfor gerne give et nuanceret indblik i fænomenet p-hacking, som er en meget aktuel og debatteret problemstilling lige nu.

2 Metode

P-hacking kan være enormt svært at opdage, da det kræver en vis gennemsigtighed i forskernes arbejdsproces. Vi har altså brug for indsigt i de valg, der er truffet hen af vejen for at kunne klassificere det som p-hacking. Dette gør p-hacking til en meget kompleks størrelse, som gør det svært at analysere ved kvantitative undersøgelser. Der findes nemlig ingen grundige sociologiske undersøgelser af fænomenet, men blot eksempler på dårlig akademisk praksis. Vores fælles udforskning af p-hacking befinder sig i en tidlig og eksplorativ fase, og derfor er det vigtigt, at vi får en dybdegående og kvalitativ indsigt i dette fænomen.

Derfor vil denne opgave tage udgangspunkt i forskellige cases, som kan danne fundamentet for en analyse, der viser forskellige aspekter af p-hacking - og som gerne skulle give en dybere forståelse af de underliggende årsager, metoder og konsekvenser af netop denne praksis.

I artiklen "Five Misunderstandigs About Case Studies" beskriver Flyvbjerg (2006) nogle strategier, man kan overveje i forbindelse med udvælgelsen af cases. Han skriver blandt andet, at hvis man ønsker at tilegne sig meget viden om et problem eller fænomen, er en repræsentativ eller random udvælgelse ikke altid at foretrække - for som han også skriver:

"Atypical or extreme cases often reveal more information because they activate more actors and more basic mechanism in the situation studied" (Flyvbjerg, 2006)

De tre cases kunne ikke vælges repræsentativt - så især to cases er ekstreme (og meget kendte) tilfælde, der indeholder adskillige aspekter og detaljer, som senere kan give anledning til en diskussion af flere problemstillinger. Desuden vil jeg prøve at trække de vigtigste pointer ud af de forskellige sager, hvorfor nogle cases vil blive gennemgået mere detaljeret end andre.

Derudover har der i udvælgelsen af cases været fokus på følgende:

- **Motiv:** De tre cases er blandt andet udvalgt med henblik på at illustrere både bevidst og ubevidst p-hacking, hvor især Case I og Case III er hinandens modpoler.
- **Metode:** De forskellige cases repræsenterer også forskellige metoder inden for p-hacking, selvom der uden tvivl findes mange flere.
- **Efterspil:** De tre cases illustrerer desuden, at der kan være vidt forskellige efterspil, som giver stof til en diskussion omkring konsekvenser og ansvar.

Casestudier kan give et nuanceret billede af en kompliceret problemstilling - og netop dette er fordelagtigt, når man skal analysere et begreb som p-hacking. Derfor vil en undersøgelse af disse cases ikke give et detaljeret og præcist overblik over omfanget, men forhåbentlig afdække de mere subtile og ofte oversete valg, der bidrager til fremkomsten af p-hacking.

"It is often more important to clarify the deeper causes behind a given problem and its consequences than to describe the symptoms of the problem and how frequently they occur" (Flyvbjerg, 2006)

3 Teori

For at forstå begrebet P-hacking, er det essentielt også at forstå den bagvedliggende teori. Derfor vil dette afsnit uddybe begreber som hypotesetest, teststørrelser, signifikansniveau, p-værdi og afslutningsvis p-hacking.

3.1 Nulhypotesen

I langt de fleste videnskabelige forsøg arbejder man med hypoteser. En hypotese er i virkeligheden et videnskabeligt gæt, som er selve byggestenen for en senere analyse. Hypoteserne trækkes dog ikke tilfældigt i en tombola, men er ofte baseret på observationer, teorier eller allerede eksisterende viden, som man ønsker at teste gennem eksperimentelle eller teoretiske metoder.

"Statistikerens hypotese siger altid, at verden er simpelt indrettet - i den forstand at den kan beskrives med få parametre. Vi vil tro på, at verden er simpel, medmindre det viser sig at være i modstrid med eksperimentelle kendsgerninger." (Hansen, 2012)

Statistikere benytter sig nemlig ofte af en *nulhypotese*. Nulhypotesen, som oftest betegnes H_0, er en antagelse om, at der ikke er nogen signifikant sammenhæng mellem de undersøgte variable. Denne hypotese antager derfor (med reference til ovenstående citat), at verden er simpel. Herefter kan man så forkaste hypotesen, hvis en senere analyse rent faktisk viser, at verden (med et vist signifikansniveau) er mere kompleks.

Foruden nulhypotesen opstilles en *alternativ hypotese* h_1, som i ordets forstand er et alternativ til nulhypotesen - dvs. den hypotese, som kunne gælde, hvis man ender med at afvise nulhypotesen.

Lad os starte med et banalt eksempel, hvor man fx. ønsker at kigge på højden for hhv. mænd og kvinder - og om der er en korrelation mellem variablerne køn og højde. Nulhypotesen vil i dette eksempel være, at der *ikke* er en forskel på den gennemsnitlige højde blandt mænd og kvinder, altså:

$$H_0 : \quad \mu_k = \mu_m$$

hvor μ_k og μ_m er middelværdien for hhv. kvinder og mænds højde fra vores fiktive data.

Tilsvarende kunne en alternativ hypotese være en af følgende muligheder:

$$\begin{aligned} H_1 : &\quad \mu_k \neq \mu_m \\ H_1 : &\quad \mu_k < \mu_m \\ H_1 : &\quad \mu_k > \mu_m \end{aligned}$$

Spørgsmålene er nu - kan vi afvise nulhypotesen? Hvordan kan vi afvise den? Og hvornår kan vi afvise den?

For at svare på disse spørgsmål, skal vi kigge nærmere på begreberne *teststørrelse, signifikansniveau og p-værdier.*

3.2 Teststørrelser, signifikansniveau og p-værdi

For at forstå begreberne, inddrages nu to grupper af stokastiske variable, nemlig $X_1, \dots X_n$ og $Y_1, \dots, Y_m$, som repræsenterer to uafhængige stikprøver fra to forskellige grupper - fx. målinger af mænd og kvinders højde.

Disse variable følger fordelingerne P og Q, således at $X_i(\mathbb{P}) = P$ og $Y_i(\mathbb{Q}) = Q$

Nulhypotesen vil da lyde: $H_0 : Q = P$, som altså er hypotesen om, at de to fordelinger er ens.

I lærebogen "The Mathematics Behind ModernDive" (Tolver og Hansen, 2024) beskrives teststørrelsen d som en funktion $d: \mathcal{X}^{n+m} \to \mathbb{R}$, der måler afvigelsen fra nulhypotesen. Den transformerer altså vores stikprøvedata fra de to stokastiske variable til et enkelt tal. Dette tal afspejler den *observerede* afvigelse mellem de to stikprøver - og kan skrives som $D = d(X_1, \dots, X_n, Y_1, \dots, Y_m)$.

Dette betyder altså, at teststørrelsen D måler, hvor meget de observerede data afviger fra det, der ville forventes, hvis nulhypotesen var sand. Det er derfor et vigtigt element, når man skal teste en hypotese, da den siger noget om, hvor ekstreme de observerede data er.

Nu kan man introducere to nye vigtige begreber, nemlig *signifikansniveau* og *p-værdi*.

Tolver og Hansen (2024) har følgende definition af signifikansniveau:

Definition 2.2.2 (Signifikansniveau) - oversat fra engelsk

Lad $c \geq 0$ og lad D være en teststørrelse for nulhypotesen $H_0 : P = Q$.

Vi kalder c en kritisk værdi for testen, hvis vi afviser H_0, når $|D| > c$.

En test af H_0 med kritisk værdi c har signifikansniveau α hvis:

$$P(|D| > c) \leq \alpha$$

under H_0.

Vi ser altså, at den kritiske værdi c er en øvre grænse, som teststørrelsen D (under nulhypotesen) skal overskride før man kan afvise nulhypotesen.

$P(|D| > c)$ er sandsynligheden for, at vores teststørrelse under nulhypotesen overskrider vores kritiske værdi. Den er et mål for, hvor stor risikoen er for at afvise nulhypotesen, selvom den faktisk er sand. Denne fejl kaldes en type-1 fejl, og dette vil blive beskrevet nærmere i afsnit 3.3 om fejltyper. Ofte beslutter man på forhånd, hvor stor denne risiko må være ved at fastlægge en grænse. Denne værdi kaldes signifikansniveauet, der ofte betegnes α.

I langt de fleste tilfælde vælges et signifikansniveau på 0.05, hvilket betyder at man kun forkaster nulhypotesen, hvis risikoen for at tage fejl er under 5 %.

P-værdi:
Nu kan vi introducere *p-værdien* - som er sandsynligheden for at observere en teststørrelse under nulhypotesen, der er mere ekstrem end den faktisk observerede teststørrelse.

$$\mathbb{P}(|D| > |d_{n,m}|)$$

hvor $d_{n,m}$ er den observerede værdi af teststørrelsen baseret på vores stikprøvedata. Der er altså valgt en kritisk værdi som værende den observerede teststørrelse fra vores stikprøvedata.

Hvis denne p-værdi er mindre end et forudbestemt signifikansniveau α, afviser vi nulhypotesen til fordel for den alternative hypotese - altså hypotesen om, at der er en forskel mellem de to gruppers fordelinger.

3.3 Fejltyper

Man støder ofte på sætningen *"... og derfor kan man kan ikke forkaste nulhypotesen"*, og sætningen er såmænd også blevet brugt i denne opgave. Al statistik beror sig på en vis usikkerhed - det er klart. Derfor kan man selvfølgelig heller ikke konkludere, at en nulhypotese er sand, da vi netop tidligere har kigget på et begreb som signifikansniveau.

- men usikkerhed kan ikke eksistere uden fejl.

Fejl opstår til tider, og dem kan man kategorisere. I nedenstående tabel ses en oversigt over de mulige udfald, der er, når man tester en nulhypotese - herunder de fejl, som kan opstå.

H_0	SAND	FALSK
Acceptere (Ej forkaste)	✓	Type 2-fejl
Forkaste	Type 1-fejl	✓

SAND og FALSK repræsenterer den universelle sandhed om nulhypotesen, hvor Acceptere (Ej forkaste) og Forkaste er den beslutning, man tager, baseret på testen.

Type 1-fejl: Man forkaster en sand nulhypotese (falsk positiv)
Type 2-fejl: Man accepterer en falsk nulhypotese (falsk negativ)

Sandsynligheden for hvert udfald ses ligeledes i nedenstående tabel.

H_0	SAND	FALSK
Acceptere (Ej forkaste)	$1-\alpha$	β
Forkaste	α	$1-\beta$

Som beskrevet tidligere repræsenterer α signifikansniveauet, som altså er sandsynligheden for at lave en type 1-fejl. For at minimere denne type fejl kan man fx. sænke signifikans-nivaeuet, men hvis man accepterer flere hyposer, vil man også acceptere flere "falske" hypoteser - altså type 2-fejl, og vice versa.

Vender vi tilbage til det banale eksempel med mænd og kvinders højde, er det måske svært at argumentere for, at den ene fejl er værre end den anden. Piben får dog en lidt anden lyd, hvis vi fx kigger mod medicinalindustrien eller ved diagnosticering af sygdomme. Langt de fleste vil nok hellere diagnosticeres fejlagtigt end at blive raskmeldt, hvor man i virkeligheden er syg.

I disse og mange andre tilfælde kan en type-1 fejl have store konsekvenser.

En af de bedste (hvis ikke *den* bedste) metode til at undgå en type 1-fejl er at undgå **p-hacking**.

3.4 P-hacking

P-hacking kan, som i ordets bogstavligste forstand, tolkes som "hacking af p-værdi". Begrebet indbefatter en usund praksis, hvor man bevidst eller ubevidst kan manipulere med p-værdien, hvis man ønsker at publicere et bestemt resultat. P-hacking er ret problematisk, idet det øger risikoen for type 1-fejl, hvor man fejlagtigt konkluderer en sammenhæng, som faktisk ikke er der.

Begrebet P-hacking blev første gang introduceret i artiklen "P -Curve: A Key to the File-Drawer" af psykologerne Uri Simonsohn, Leif D. Nelson og Joseph P. Simmons.

"While collecting and analyzing data, researchers have many decisions to make, including whether to collect more data, which outliers to exclude, which measure(s) to analyze, which covariates to use, and so on. If these decisions are not made in advance but rather are made as the data are being analyzed, then researchers may make them in ways that self-servingly increase their odds of publishing. Thus, rather than placing entire studies in the file-drawer, researchers may file merely the subsets of analyses that produce nonsignificant results. We refer to such behavior as ***p-hacking***" (Simonsohn et al., 2014)

Som beskrevet i ovenstående citat, skal forskere træffe mange valg i forbindelse med deres forskning, og det er afgørende, at man træffer disse beslutninger *inden* undersøgelsen er gået i gang.
Simonsohn et al. (2014) beskriver, hvordan man bl.a. kan ekskludere outliers, manipulere med variable osv. Disse valg kan bevidst eller ubevidst påvirke p-værdien, og dermed føre til p-hacking. Det kan bl.a. forekomme ved følgende metoder, men der findes sikkert mange flere:

- **Sample Size**
  Man afslutter en dataindsamling, når et signifikant resultat opnås - uden at tage hensyn til en forudbestemt samplesize. Man fortsætter altså med at indsamle data til man har den ønskede p-værdi og stopper herefter.
- **Fjerne outliers**:
  Man fjerner outliers fra datasættet for at opnå et signifikant resultat - og altså ikke af teoretiske grunde. Her kunne der være tale om outliers, som øgede p-værdien og som dermed ville være fordelagtigt at fjerne.
- **Manipulation med variabler:**
  Man manipulerer med variablerne eller ændrer på, hvilke betingelser, der inkluderes i analysen - fx ved at ændre responsvariablen, hvis det ville resultere i en mindre p-værdi.
- **Overdreven hypotesetestning:**
  Man udfører rigtig mange hypotesetests, hvilket øger sandsynligheden for, at udfaldet er tilfældigt.
- **Modeltilpasning:**
  Man prøver mange forskellige modeller af indtil man opnår det ønskede resultat og/eller forskellige teststørrelser.
- **Selektiv rapportering af resultater:**
  Man rapporterer kun statistisk signifikante resultater, mens ikke-signifikante resultater og beskrivelser om ændringer i studiet udelades.

P-hacking er altså en manipulation af p-værdien, hvilket betyder at der fejlagtigt kan drages en forkert beslutning på baggrund af denne værdi. Dette betyder, at p-hacking kan medføre type-1 fejl - altså et falsk-positivt resultat, hvor man forkaster en sand nulhypotese.

I en artikel i "Psychological Science" udfører Simonsohn et al. (2011) en række simuleringer for at illustrere, hvordan forskellige "researcher degrees of freedom" – valgene, som forskere træffer under dataindsamling og analyse – kan påvirke sandsynligheden for falsk-positive resultater.

Simulationen var opbygget ud fra situationer, som er klassiske p-hacking metoder:

- **Situation A:** Man må vælge mellem forskellige afhængige variabler
- **Situation B:** Man må fortsætte med at indsamle data indtil signifikante resultater opnås.
- **Situation C:** Man må vælge at kontrollere for visse kovariater.
- **Situation D:** Man rapporterer kun resultater fra visse betingelser, mens man udelader andre.

Resultatet af de 15000 simulationer kan ses i nedenstående tabel, og udfaldet kan virke overraskende. Tabellen viser nemlig, at selv en lille fleksibilitet i forskernes valg, kan øge sandsynligheden for at opnå et falsk-positivt resultat markant. Det ses fx. at hvis forskere kan vælge mellem to afhængige variabler (Situation A) er sandsynligheden 9.5 %, hvilket er næsten dobbelt så stor en sandsynlighed, som den normale tærskel på 5 %. Simonsohn et al. (2011) viser desuden i tabellen, at sandsynligheden kan stige helt op til ca. 60 %, når de kombinerede flere af disse frihedsgrader.

Table 1. Likelihood of Obtaining a False-Positive Result

	Significance level		
Researcher degrees of freedom	$p < .1$	$p < .05$	$p < .01$
Situation A: two dependent variables ($r = .50$)	17.8%	9.5%	2.2%
Situation B: addition of 10 more observations per cell	14.5%	7.7%	1.6%
Situation C: controlling for gender or interaction of gender with treatment	21.6%	11.7%	2.7%
Situation D: dropping (or not dropping) one of three conditions	23.2%	12.6%	2.8%
Combine Situations A and B	26.0%	14.4%	3.3%
Combine Situations A, B, and C	50.9%	30.9%	8.4%
Combine Situations A, B, C, and D	81.5%	60.7%	21.5%

Tabel 1: Tabellen viser en simulation på 15000 samples, hvor man afprøver større fleksibilitet i forskeres frihedsgrader - og som viser sandsynligheden for at få en p-værdi under hhv. 0.1, 0.05 og 0.01 Simonsohn et al. (2011)

Denne tabel cementerer derfor, at den sande risiko for at opnå et falsk-positivt resultat i praksis kan være meget højere end de 5 % - og at omfanget af type-1 fejl derfor kan være større end man regner med.

4 Analyse

4.1 Case I: Pizza buffet

En af de mest kendte tilfælde af p-hacking involverer Brian Wansink, som er en amerikansk professor, der var leder af Cornell Universitys Food and Brand Lab. Han opnåede bred anerkendelse og en del offentlig opmærksomhed for sin forskning i menneskers spisevaner. Han formidlede bl.a. sin forskning i populære tv-shows som *Oprah*, samt skrev bestselleren "Mindless Eating: Why We Eat More Than We Think" - hvor han præsenterede sin forskning og gav praktiske råd til, hvordan man kunne ændre sine spisevaner. Wansink optrådte altså på den helt store scene, hvor han blev kaldt "The Sherlock Holmes of Food" (Bartlett, 2017). Et tilnavn, som anerkendte hans store detektivarbejde inden for menneskelig adfærd og spisevaner. I dag bliver han dog ikke længere associeret med englands mest berømte detektiv, da hans forsknings troværdighed er blevet svækket markant. Begyndelsen til enden skete i november 2016, da han udgav et nu meget berømt blogindlæg på sin hjemmeside med titlen "The Grad Student Who Never Said No".

I dette blogindlæg fortæller han om en ph.d. studerende fra Tyrkiet, som kom til hans laboratorium for at arbejde som forsker i 6 måneder. Han fortæller, at hun herefter blev tildelt et 'mislykket' datasæt indeholdende observationer fra en italiensk buffet restaurant, som ikke havde vist nogen signifikante resultater. Wansink håbede derfor, at hun kunne "redde" noget fra datasættet ved at analysere det på nye måder.

"This cost us a lot of time and our own money to collect. There's got to be something here we can salvage because it's a cool (rich & unique) data set." (Wansink, 2016)

I stedet for at anerkende studiets manglende resultater, blev den ph.d. studerende i stedet opfordret til at analysere data igen og igen - med henblik på at opnå resultater, der kunne publiceres. Data havde været tidskrævende og dyrt at indsamle, hvorfor vi må formode, at Wansink nødigt så det gå til spilde.

"Every day she came back with puzzling new results, and every day we would scratch our heads, ask "Why," and come up with another way to reanalyze the data with yet another set of plausible hypotheses. Eventually we started discovering solutions that help up regardless of how we pressure-tested them. I outlined the first paper, and she wrote it up, and every day for a month I told her how to rewrite it and she did." (Wansink, 2016)

En analyse af ovenstående citat viser tydeligt, at Brian Wansinks arbejdsmetode indeholder mange elementer, som direkte kan klassificeres som p-hacking. Sammenligner man citatet med vores eksempler på p-hacking (3.4) træder især to eksempler/metoder ud.

- **Manipulation med variabler:**
  Processen med at "finde på en anden måde at analysere data på" antyder en manipulation med variablerne, hvor forskellige responsvariable muligvis også blev overvejet ind til man fik et statistisk signifikant resultat.

- **Overdreven hypotesetestning:**
  Citatet illustrerer klart denne praksis gennem den massive og kontinuerlige hypotesetestning,

samt adskillige nye analyser baseret på nye hypoteser. Denne tilgang øger sandsynligheden for tilfældigt at finde signifikante resultater uden en forudbestemt plan for undersøgelsen.

Deres gentagne analyser af det oprindelige datasæt resulterede i hele fem publikationer, der blandt andet fremlagde, at mænd spiser mere i selskab med kvinder - og at man oftere fortryder at spise buffet, hvis prisen er lav. De fem publikationer er nedenfor listet med nummerering til senere analyse.

1. Just, David R., Ozge Sigirci, and Brian Wansink (2014), **"Lower Buffet Prices Lead to Less Taste Satisfaction"**, Journal of Sensory Studies, 29:362-370.
2. Just, David R., Ozge Siğirci, and Brian Wansink (2015), **"Peak-end Pizza: Prices Delay Evaluations of Quality"**, Journal of Product & Brand Management, 24:7, 770-778
3. Kniffin, Kevin, Ozge Sigirci and Brian Wansink (2015), **"Eating Heavily: Men Eat More in the Company of Women"**, Evolutionary Psychological Science, 1-9.
4. Siğirci, Ozge and Brian Wansink (2015), **"Low Prices and High Regret: How Pricing Influences Regret at All-You-Can-Eat Buffets"**, BMC Nutrition, 1:36, 1-5
5. Sigirci, Ozge, Marc Rockmore, and Brian Wansink (2016), **"How Traumatic Violence Permanently Changes Shopping Behavior"**, Frontiers in Psychology, 7:1298.

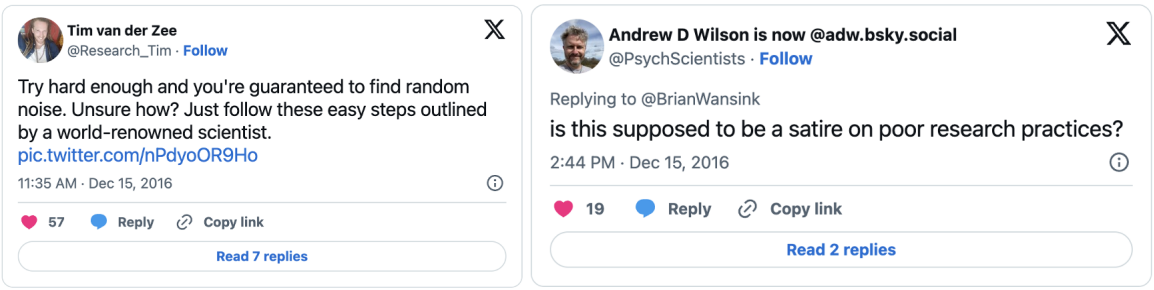
Brian Wansink ønskede i sit blogindlæg at fortælle en historie, der skulle understrege værdien i "ikke at sige nej", samt hvordan den rette indstilling og ihærdighed kan lede til stor akademisk succes. Dette stod i stærk kontrast til en anden postdoc, som havde afslået at arbejde med samme datasæt pga. tid og andre prioriteter, som fx. *"Facebook, Twitter, Game of Thrones, Starbucks, spinning class ..."* (Wansink, 2016). Det fremgår ret tydeligt, at Wansink ønsker at fremhæve denne postdoc negativt, da han oplyser, at personen kun publicerede en fjerdedel af hvad den tyrkiske kvinde formåede - og at personen i øvrigt var misundelig på hende.

"Yet most of us will never remember what we read or posted on Twitter or Facebook yesterday. In the meantime, this Turkish woman's resume will always have the five p apers" (Wansink, 2016)

Ovenstående viser, at Wansink fokuserer på antallet af publikationer - og at det faktisk er et mål for akademisk succes.
Dette fokus på publicering - og det faktum, at de kun rapporterede succeserne, henleder til endnu et eksempel på formodet p-hacking (4.3):

- **Selektiv rapportering af resultater:** Der blev publiceret fem artikler fra det oprindelige datasæt, men de undlod at rapportere alle de gange, hvor undersøgelser af datasættet ikke havde givet noget signifikant resultat.

Wansinks blogindlæg fik en hård medfart, og mange stillede sig undrende over for hans metode, der mest af alt blev kaldt en "manual til p-hacking". En læser gik skridtet videre og spurgte, hvorvidt hans indlæg skulle forstås som satire.



Undersøgelse af artikel 1-4:
Blandt de mange kritikere var Jordan Anaya, Nicholas J. L. Brown og Tim van der Zee, hvor sidstnævnte ses i ovenstående twitter-opslag. De besluttede sig for at lave en dybdegående undersøgelse af fire ud af de fem publikationer, som Wansink nævnte i sit blogindlæg. De valgte at undersøge artikel 1-4, som er listet på forrige side, da disse tog udgangspunkt i samme datasæt. I forbindelse med denne undersøgelse opdagede de en lang række fejl i Wansinks arbejde, som resulterede i artiklen: "Statistical heartburn: an attempt to digest four pizza publications from the Cornell Food and Brand Lab" (Anaya et al., 2017)

I undersøgelsen havde de ikke adgang til det oprindelige datasæt, selvom de flere gange havde henvendt sig til alle forfattere, der stod listet på de fire publikationer. Derfor benyttede de sig af statistiske værktøjer og teknikker, der kan bruges - selvom man ikke har adgang til det oprindelige datasæt. I deres analyse opdagede de omkring 150 fejl og uoverensstemmelser i Wansinks publikationer, der blandt andet indbefattede fejl i detaljeringsgrad, teststørrelser og samplesizes. Det er dog vigtigt at bemærke, at selvom de fandt mange fejl i de undersøgte artikler, er det ikke alle fejl, som nødvendigvis kan betragtes som p-hacking. Nogle af dem er blot fejl (omend ret graverende fejl) i datahåndteringen og analyseprocessen. Anaya et al. (2017) sammenlignede de rapporterede data fra de fire artikler, hvor de fx. identificerede uoverensstemmelser af middelværdier og standardafvigelser - som ikke var mulige givet den angivne sample size. Disse uoverensstemmelser blev yderligere bekræftet, da de fandt forskelle i sample size på tværs af de forskellige artikler.

Netop uoverensstemmelser af sample size kan være tegn på p-hacking.

Anaya et al. (2017) beskriver, at der i artikel 4 er en uoverensstemmelse mellem det datasæt, der bliver analyseret og den hypotese, de undersøger. Det fremgår nemlig i artiklen, at deltagerne skulle have spist mindst ét stykke pizza for at blive inkluderet i analysen. Nedenstående tabel danner grobund for artikel 4, og den viser en Likertscala, hvor der samlet set er 95 respondenter, som angiveligt skulle have spist mindst ét stykke pizza. Tabellen viser altså (ifølge antagelsen om, at alle spiste mindst ét stykke), at der tilsyneladende ikke er en eneste person blandt de 135 medvirkende, der spiste mere end 3 stykker.

Table 3

	1 Piece			2 Pieces			3 Pieces		
	$4 (*N* = 18)	$8 (*N* = 19)	*F* test	$4 (*N* = 18)	$8 (*N* = 21)	*F* test	$4 (*N* = 7)	$8 (*N* = 12)	*F* test
I ate more pizza than I should have	2.63 (2.06)	1.76 (1.82)	1.62	4.82 (2.55)	3.53 (2.39)	2.47	6.00 (2.00)	4.40 (3.24)	1.34
I feel guilty about how much I ate	2.39 (1.94)	2.26 (1.79)	0.04	3.44 (2.48)	1.68 (1.42)	7.13	3.71 (1.50)	2.90 (2.08)	0.78
I am physically uncomfortable	2.17 (1.89)	1.955 (1.68)	0.14	2.94 (2.13)	1.28 (0.46)	8.11	2.43 (1.51)	2.10 (1.91)	0.14
I overate	2.11 (1.81)	1.67 (1.28)	0.72	3.89 (2.59)	1.53 (1.02)	1.63	3.71 (1.79)	3.50 (2.95)	0.03
I ate more than I should have	2.50 (2.20)	2.00 (1.45)	0.67	4.28 (2.44)	2.05 (1.72)	10.36	4.57 (2.23)	4.00 (3.02)	0.18

Tabel 3 - Anaya et al. (2017)

Påstanden om, at de 95 medvirkende alle havde spist mindst ét stykke pizza, står dog i stærk kontrast til nedenstående tabel fra artikel 1. Denne tabel viser ret tydeligt, at samlet 122 personer havde en holdning til smagen af pizza, hvorfor man må antage, at alle 122 personer også har spist et stykke.

Table 2

	$4 buffet (*N* = 62)	$8 buffet (*N* = 60)	*F* test (*p* value)
The pizza, in general, tasted really great	6.89 (1.39)	7.44 (1.60)	4.24 (0.04)
The first piece of pizza I ate tasted really great	7.08 (1.30)	7.45 (1.60)	1.97 (0.16)
The first piece of pizza I ate was very satisfying	7.08 (1.37)	7.34 (1.70)	0.82 (0.37)
The first piece of pizza I ate was very enjoyable	7.05 (1.40)	7.47 (1.55)	2.40 (0.12)
The middle piece of pizza I ate tasted really great	6.68 (1.49)	7.97 (1.21)	15.42 (0.00)
The middle piece of pizza I ate was very satisfying	6.68 (1.49)	7.97 (1.21)	14.69 (0.00)
The middle piece of pizza I ate was very enjoyable	6.64 (1.48)	7.81 (1.22)	12.48 (0.00)
The last piece of pizza I ate tasted really great	6.15 (1.89)	7.58 (1.39)	15.16 (0.00)
The last piece of pizza I ate was very satisfying	6.16 (1.87)	7.41 (1.55)	10.99 (0.00)
The last piece of pizza I ate was very enjoyable	5.98 (1.86)	7.45 (1.52)	15.60 (0.00)

Tabel 2 - Anaya et al. (2017)

En nærmere undersøgelse af tabellerne tydede derfor på, at de 95 respondenter, som blev benyttet i artikel 4 faktisk var udvalgt, fordi de *ikke* havde spist mere end tre stykker pizza (Anaya et al., 2017). Wansink og den tyrkiske kvinde havde altså udvalgt de personer, der havde spist enten et, to eller tre stykker pizza.

Anaya et al. (2017) undrer sig derfor over dette kriterium, da det erklærede formål med studiet var at undersøge skyldfølelse, mæthedsfølelse og overspisning. De påpeger derfor, at det virker ulogisk at udelukke netop de personer, der spiste den største mængde pizza - og undrer sig i øvrigt også over, at de udelukkede andre kalorierige retter, som fx. pasta.

De italesætter en vigtig pointe - for hvorfor fravalgte man de personer, som spiste mest pizza foruden andre kalorierige retter i et studie, der handler om skyld og overspisning? Svaret er højst sandsynligt - *fordi det gav et resultat.*

Det er som tidligere beskrevet enormt svært at påvise p-hacking, da det kræver en vis gennemsigtighed i arbejdsprocessen. Det er derfor svært at konkludere, hvilke metoder der er blevet brugt til at få et signifikant resultat i artikel 1-4. Det er dog et statistisk faktum, at man kan øge chancen for at opnå et signifikant resultat, hvis man selektivt inkluderer eller udelukker data fra undersøgelsen - og netop dette henviser til endnu to eksempler fra teori (3.4)

- **Sample Size**
  Der er tale om p-hacking, hvis arbejdsprocessen indebar selektive tests. Fx ved først at analysere effekten af at spise ét pizzastykke, derefter to, og så tre - og processen så afbrydes, så snart p-værdien er under det ønskede signifikansniveau.

- **Outliers**
  Det er tale om p-hacking, hvis man fjerner deltagere, som i dette tilfælde spiste mere pizza end gennemsnittet, fordi de øger p-værdien - især hvis man gør det uden en solid teoretisk begrundelse.

Der tegner sig altså et billede af, at Brian Wansink havde en aldeles sjusket tilgang til videnskabelig forskning, da samtlige af de metoder til p-hacking, som er opsummeret i teoriafsnittet, kan sættes i forbindelse med hans arbejde.

"None of us can remember encountering a set of articles with as many inconsistencies and unresolved questions in the basic reporting of results as in this case." (Anaya et al., 2017)

Wansinks blogindlæg satte selvfølgelig gang i en lang række undersøgelser af hans arbejde. Undersøgelserne afslørede så alvorlige problemer, at flere af hans udgivelser blev trukket tilbage, og han endte med at forlade Cornell University.

Brian Wansink var tilsyneladende helt uvidende omkring sin p-hacking - og det tyder på, at hans primære motiv var at tilføje flere publikationer til sin i forvejen lange liste. Det formåede han også at gøre, men det blev på bekostning af en videnskabelig integritet.

Sagen rejser selvfølgelig en diskussion omkring de incitamenter, der driver videnskabelig forskning. Hvis motivationen i højere grad handler om antallet af publikationer fremfor et ønske om at gøre verden mere oplyst - har videnskaben så i virkeligheden mistet sit formål? Og når en så anerkendt og højt profileret forsker (som Brian Wansink) bliver afsløret ved et tilfælde - hvilke konsekvenser har det så for videnskabens generelle troværdighed?

Disse spørgsmål vil blive taget op i en senere diskussion.

4.2 Case II: Motorcykler og fuldmåne

Forestil dig en mørk aften, hvor du kommer kørende på din motorcykel. Du ser pludselig fuldmånen, der lyser kraftigt op på himlen. Du betragter fuldmånen... men lidt for længe, for du mister koncentrationen. Du mister faktisk også så meget kontrol over din motorcykel, at du ender i et uheld med dødelig udgang.

Det er barskt, men ikke desto mindre et realistisk scenarie ifølge Donald A. Redelmeier og Eldar Shafir, som i 2017 publicerede artiklen "The full moon and motorcycle related mortality: population based double control study" (Redelmeier og Shafir, 2017). I denne artikel undersøgte de hypotesen om, at antallet af dødsfald i forbindelse med motorcykelulykker påvirkes af, hvorvidt der er fuldmåne eller ej. Forskningen er baseret på en omfattende gennemgang af 13.029 dødsfald i motorcykelulykker i USA fra 1975 til 2014, hvor dødsfaldene blev inddelt i en fuldmånegruppe og en kontrolgruppe.

"We hypothesized that because people's attention is naturally drawn to a full moon, it might contribute to fatal motorcycle crashes. In particular, glancing at a full moon takes the motorcyclist's gaze off the road, which could result in a loss of control." (Redelmeier og Shafir, 2017)

Konklusionen var, at der faktisk var en øget risiko for dødsfald under fuldmåne med en relativ risiko (RR) på 1.05 og en stigning på 226 yderligere dødsfald gennem hele perioden. Dvs. at risikoen for at være involveret i en motorcykelulykke, hvor en person dør, er 5 % højere på nætter med fuldmåne sammenlignet med nætter uden fuldmåne.

Men - inden vi nu for alvor begynder at betragte en motorcykel under fuldmåne som en sort kat, der går over vejen, vil det give mening først at dykke ned i de beslutninger, som Redelmeier og Shafir (2017) traf undervejs i deres arbejdsproces. Det viser sig nemlig, at nogle af disse beslutninger måske kunne være truffet anderledes. Valgene indbefattede blandt andet:

1. Inkluderede kun motorcykelulykker
2. Tidsintervallet for "night time" blev sat fra kl. 16 til 8 om morgenen.
3. Fuldmånendatoerne var baseret på tidzonen i London.
4. Kontrolgruppen var ulykker, som skete syv dage før og efter en fuldmåne.

Test af hypotese:

Hypotesen forsøgte Smith (2023) at undersøge igen. Han valgte dog at træffe nogle andre valg. Først og fremmest valgte han foruden motorcykler, også at teste for alle slags køretøjer. Derudover ændrede han tidszonen, så datoerne for fuldmåne dækkede en tidszone i USA, hvor data var registreret - samt indsnævrede tidsintervallet, så "night time" nu var fra kl. 20 om aftenen til kl. 4 om morgenen. Kontrolgruppen blev nu valgt til 14 dage før fuldmåne.

Table 7.1 *Number of Accidents and Fatalities on Full-Moon Nights*

	4 p.m. to 8 a.m. window		8 p.m. to 4 a.m. window	
	All Vehicles	Motorcycles	All Vehicles	Motorcycles
Number of fatal accidents				
Full moon	37,964	**3,706**	21,158	1,985
+/- 7 days	38,075.5	**3,603.0**	20,979.5	1,974.5
+/- 14 days	38,047.0	3,669.5	21,272.0	1,997.0
Number of fatalities				
Full moon	42,222	3,873	23,637	2,080
+/- 7 days	42,414.5	3,768.0	23,433.0	2,073.5
+/- 14 days	42,462.5	3,828.5	23,834.0	2,089.5

Tabel 7.1: Tabellen viser antallet af motorcykelulykker og andre køretøjer med og uden dødelig udgang i en ny analyse. Antallet er inddelt i henholdsvis tidsrum, fuldmåne og kontrolgrupper - Smith (2023)

Ovenstående tabel viser resultatet. Tallene, som er markeret med fed, repræsenterer det nye resultat af det oprindelige setup - nemlig målinger for motorcykeluheld på fuldmåne og tilsvarende $\pm$ 7 dage, dog beregnet ud fra den nye tidszone i USA. Disse tal viser, at der stadig var flere motorcykeluheld ved fuldmåne, men at det kun svarer til en øget risiko på omkring 2.9 % - hvilket er en del under de oprindelige 5 %. Tilsvarende respræsenterer tallene i højre side det nye tidsinterval. Der var 1.985 uheld (med dødelig udgang) under fuldmåne og 1.997 uheld omkring nymåne - dvs. færre uheld opstod under fuldmåne.

"Of the sixteen comparisons in Table 7.1, the number of accidents or fatalities were higher on full-moon nights in eight cases and lower in eight cases ... Here, the evidence that it is unusually dangerous to drive on full-moon nights is not compelling." (Smith, 2023)

Studiet om motorcykeluheld under fuldmåne illustrerer derfor meget tydeligt, hvor afgørende de valg, man træffer i forskningsprocessen, kan være - men også at forskellige akkumuleringer af valg leder ud på forskellige stier og dermed forskellige destinationer. Der tegner sig et billede af, at Redelmeier og Shafir (2017) med deres valg slentrede ubekymret ned af en meget specifik sti, som førte til et statistisk signifikant resultat. Når man gør dette - og hverken undersøger eller italesætter andre mulige stier, der ville have givet andre resultater - er der tale om *p-hacking*.

Problemet er, at vi ikke ved, hvorvidt der er tale om utilsigtet eller bevidst p-hacking - og om de valgte den statistisk signifikante sti ved et tilfælde eller faktisk havde undersøgt mange andre stier inden. Det viser blot, at forskellige valg kan give forskellige p-værdier, men det beviser ikke nødvendigvis en bevidst manipulation af data. Dette er præcist problemet med p-hacking - det er *meget* svært at påvise.

Artiklen er stadig at finde på hjemmesiden, og der er hverken disclaimers, kommentarer eller andre indikationer, som giver indsigt i dens svagheder. Det er derfor oplagt at diskutere, hvorvidt tidskrifterne har et ansvar i at sikre en videnskabelig integritet. Derudover er det også essentielt at undersøge, hvilke redskaber der skal til, hvis vi skal kunne filtrere artikler som denne fra.

4.3 Case III: Study 329 - Paroxetin til unge

I den lidt mere alvorlige ende af p-hacking-skalaen, findes sagen om "Study 329". Et studie, som havde vidtrækkende konsekvenser for helt almindelige mennesker, og som blev katalysatoren for en stor diskussion om gennemsigtighed i medicinalindustrien.

Study 329 var et forskningsprojekt fra USA, der i 1990'erne undersøgte virkningen af det antidepressive lægemiddel Paroxetin på unge under 18 år med depression. Forskningsprojektet var finansieret af GlaxoSmithKline (GSK), som er en medicinalvirksomhed, der bla. sælger paroxetin i pilleform under navnet Paxil. Undersøgelsen af de unge med depression resulterede i en forskningsartikel (Keller, 2001), der blev udgivet i tidskriftet "American Academy of Child and Adolescent Psychiatry (JAACAP)". Artiklen tog udgangspunkt i kliniske forsøg på 275 unge med depression, og sammenlignede altså effekten af Paroxetin med placebo. Konklusionen var kort og præcis:

"The findings of this study provide evidence of the efficacy and safety of the SSRI, paroxetine, in the treatment of adolescent depression." (Keller, 2001)

Problemet var bare, at resultaterne var vildledende. I nogle interne dokumenter, der blev offentliggjort i forbindelse med en retssag mod GSK, viste det sig nemlig, at de havde p-hacket resultaterne ved hjælp af blandt andet 'selektiv rapportering' og 'manipulation af variabler' (3.4).

Selektiv rapportering:

I artiklen "Clinical trials and drug promotion: Selective reporting of study 329" adresserer Jureidini et al. (2008) problemet med selektiv rapportering og variabelmanipulation og beskriver i detaljer, hvordan Study 329 gjorde brug af netop disse (uhensigtsmæssige) metoder.

I forsøget, der skulle undersøge effekten af paroxetin på unge, var der oprindeligt to primære mål/responsvariable. Keller (2001) ønskede først og fremmest at kigge på den samlede ændring for hver deltager på "Hamilton Depression Rating Scale (HAM-D)", som er en skala, der kan vurdere sværhedsgraden af eventuelle depressionssymptomer. Derudover ønskede de at undersøge andelen af deltagere, som reagerede på behandlingen. Det blev betragtet som en reaktion, hvis deltagernes HAM-D-måling enten var faldet til under 8 eller var reduceret med 50 % eller derover. Dette ville nemlig indikere en betydelig forbedring af deres depressionssymptomer.

Umiddelbart virker det til at være fornuftige responsvariable, hvis man vil undersøge en nulhypotese, der siger, at Paroxetin ingen virkning har på unge. Problemet for Keller (2001) var dog, at de ikke kunne vise noget som helst signifikant ud fra disse. Derfor besluttede forskerne at introducere nye responsvariable, som ikke var rapporteret i det oprindelige setup, men som tilgengæld var statistisk signifikante. På denne måde blev der tegnet et mere positivt billede af paroxetins effekt på unge med depression.

Ifølge Jureidini et al. (2008) var fire ud af de otte "ubrugelige" responsvariable blevet erstattet med fire nye, som gav et positivt resultat - og der var i øvrigt mange andre variable, der var blevet testet i løbet af analysen, som også havde vist sig som værende ikke-signifikante. Disse resultater var heller ikke blevet rapporteret.

Table 1
Outcome measures (significant results in **bold**); ordering of outcome measures is from originals

Protocol (1993, 1996) [12]	p	Final paper (2001) [5]	p
*Change in HAM-D total score	0.13	**HAM-D $\leqslant 8$**	**0.02**
*Responders (HAM-D $\leqslant 8$ or reduced by $\geqslant$50%)	0.11	*Responders (HAM-D $\leqslant 8$ or reduced by $\geqslant$50%)	0.11
Depression scale of K-SADS-L	0.07	**HAM-D depressed mood item**	**0.001**
Mean Clinical Global Improvement (CGI) score	0.09	**K-SADS-L depressed mood item**	**0.05**
Autonomous function checklist	0.15	**CGI 1 or 2**	**0.02**
Self-perception profile	0.54	Depression scale of K-SADS-L	0.07
Sickness impact scale	0.46	Mean CGI	0.09
Relapse during maintenance	0.24**	*HAM-D total score	0.13

*Protocol specified primary outcomes. **Not published, calculated by us, trend favours placebo.

Table 1 - Tabellen viser en sammenligning af responsvariablerne fra den oprindelige studieprotokol og de responsvariabler, som blev offentliggjort i den endelige artikel - hvor de tilhørende p-værdier er angivet Jureidini et al. (2008)

I ovenstående tabel ses en sammenligning af de oprindelige responsvariable og de responsvariable, der optrådte i den endelige artikel. Det fremgår heraf, at ingen af de udregnede p-værdier i det oprindelige setup var under signifikansniveauet. I højre kolonne er der dog tilføjet nye mål, hvor dem markeret med fed, viste et resultat. Det ses bla. at deres nye responsvariabel nu var HAM-D < 8 - og altså ikke længere den samlede ændring i HAM-D-målingerne.

Det står derfor ret klart, at Keller (2001) - foruden en mangelfuld rapportering omkring mislykkede forsøg - også har manipuleret med variablerne. GSK kunne nu markedsføre paroxetin som værende både effektivt og sikkert for behandling af depression hos unge - *på trods* af deres mange fejlagtige forsøg på at vise netop dette. I artiklen "Restoring study 329" (Noury et al., 2015) blev studiet analyseret på ny og gransket for fejl. Ved at anvende det oprindelige data fra Study 329, udarbejdede de en korrektion til studiet med en ny og uafhængig analyse af effekten af paroxetin. Konklusionen var dog markant anderledes:

- Paroxetin var *ikke* effektivt til behandling af de medvirkende.
- Paroxetin kunne bidrage til øget selvskade og selvmordstanker blandt de medvirkende.

Table 6 | Numbers of patients with suicidal and self injurious behaviours in Study 329 with different safety methods

	Paroxetine (n=93)	Imipramine (n=95)	Placebo (n=87)
Keller and colleagues*	5	3	1
SKB acute from CSR*	7	3	1
RIAT acute and taper from CSR	11	4 (3 definite, 1 possible)	2 (1 definite, 1 possible)

*Keller and colleagues and CSR mostly reported suicide related events as "emotional lability."

Table 6 - Tabellen er en oversigt over forskellige rapporteringer af selvskadende adfærd - Noury et al. (2015)

Tabellen viser antallet af patienter med selvskadende adfærd i Study 329, og den sammenligner de forskellige rapporteringer af disse data. I den oprindelige artikel angav Keller (2001) 5 tilfælde af

selvmordsrelateret adfærd for paroxetin og 1 for placebo. I den kliniske rapport af SKB (nu GSK), var der angivet 7 tilfælde i paroxetin-gruppen og 1 i placebo-gruppen. Den nye analyse af Noury et al. (2015) (RIAT) inkluderede data fra både den akutte fase og nedtrapningsfasen fra den kliniske rapport - og denne viste 11 tilfælde i paroxetin-gruppen og 2 tilfælde i placebo-gruppen. Disse bivirkninger - som jo går igen i alle tre rapporter, var ikke tydeliggjort i den oprindelige konklusion, men blot skjult lidt af vejen i en tabel. I den nye analyse, valgte de også at inkludere data fra udfasningsperioden, hvilket jo er et væsentligt perspektiv, når man skal undersøge effekten af paroxetin - og dets sikkerhed. Keller (2001) havde i øvrigt en tendens til at beskrive disse hændelser som "emotionel labilitet" altså følelsesmæssig ustabilitet - hvilket må siges at være en ret nedtonet måde at fortolke selvskadende adfærd (herunder selvmord) på.

Ovenstående pointer vidner om, at Studie 329 blev p-hacket. Vi så fx eksempler på 'manipulation af variabler' og 'selektiv rapportering' (teori - 3.4) - og i virkeligheden kan vi ikke udelukke, at de også har benyttet andre metoder. Spørgsmålet er så, hvorvidt der, ligesom Brian Wansink, er tale om en ubevidst samt sjusket tilgang til dataanalyse og rapportering - eller om p-hackingen foregik med fuldt overlæg?

Svaret på dette står klart - og faktisk *foruroligende* klart, når man læser et internt dokument fra GSK (Unknown, 1998), som blev offentliggjort senere i forbindelse med en retssag. Dokumentet er skrevet tre år *inden* Keller (2001) skrev den positive konklusion. Modsat denne konklusion, anerkender de i dokumentet, at Study 329 ikke viste den ønskede effekt med paroxetin til unge - og kommer så med følgende forslag:

PROPOSALS

- **Based on the current data from Studies 377 and 329, and following consultation with SB country regulatory and marketing groups, no regulatory submissions will be made to obtain either efficacy or safety statements relating to adolescent depression at this time. However data (especially safety data) from these studies may be included in any future regulatory submissions, provided that we are able to go on and generate robust, approvable efficacy data. The rationale for not attempting to obtain a safety statement at this time is as follows;**

  i) regulatory agencies would not approve a statement indicating that there are no safety issues in adolescents, as this could be seen as promoting off-label use

  ii) it would be commercially unacceptable to include a statement that efficacy had not been demonstrated, as this would undermine the profile of paroxetine.

"Position piece on the fase III clinical studies" - (Unknown, 1998)

Der var altså en antagelse om, at myndighederne ikke ville godkende en anprisning om sikkerhed, fordi de tilgængelige data ikke underbyggede en sikker brug af paroxetin hos unge - og at det (oversat til dansk) *"ville være kommercielt uacceptabelt at anerkende den manglende effekt af paroxetin, da det kunne underminere paroxetins omdømme"* .

Citatet kan sagtens stå alene, da det netop er et helt eksplicit udtryk for en stor interessekonflikt hos GSK - for efter Keller (2001) udgav artiklen steg salget af Paxil til unge markant. Dette

leder til en diskussion om integritet i forskning - for giver det mening at rådføre sig hos en marketingafdeling, som ønsker at sælge et produkt, omkring et negativt resultat af selvsamme produkt? Study 329 er derfor et godt eksempel på, hvordan kommercielle interesser risikerer at tilsidesætte videnskabelig integritet. Disse interessekonflikter, samt andre former for bias vil senere blive diskuteret yderligere.

Efter afsløringerne af Study 329 stod GSK over for nogle ret store juridiske og finansielle tæsk. GSK indgik et forlig med de amerikanske myndigheder og endte med at betale en erstatning på 3 milliarder dollars.

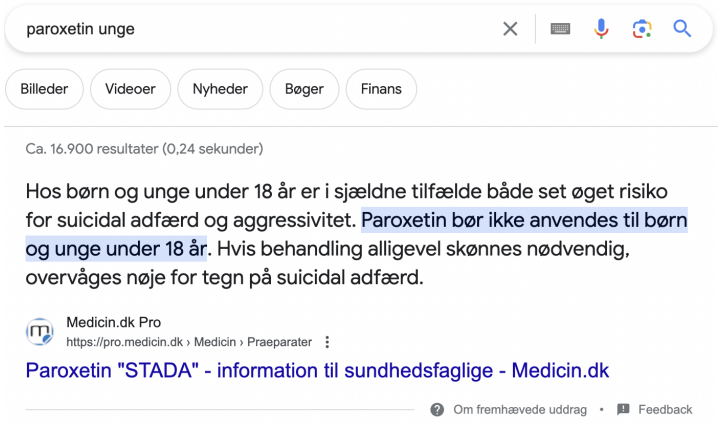
Med medicinalindustriens største erstatning nogensinde - nye analyser og interne dokumenter, der viser en manglende effekt af Paroxetin på unge - samt en dokumenteret kommerciel interesse fra GSK, skulle man tro, at der efterhånden var hamret nok søm i Study 329's kiste.

Men nej. Artiklen er den dag i dag fortsat ikke trukket tilbage af tidskriftet JAACAP - på trods af mange anmodninger om netop dette. I artiklen "Rules of retraction" (Newman, 2010) beskrives to forskeres kamp for netop at få trukket artiklen tilbage, men det er ikke lykkedes. JAACAP fastholder, at den oprindelige artikel om Study 329 ikke indeholder fejl - og at de negative resultater er inkluderet i en resultattabel, hvorfor der ikke er grundlag for at trække artiklen tilbage.

"I've been surprised how hard it's been to get editors to take action to improve the quality of their journals. They prefer to turn a blind eye." (Newman, 2010)

Denne modvilje, der bliver italesat i ovenstående citat, stiller altså spørgsmålstegn ved, hvorvidt tidskrifter lever op til deres ansvar med at sikre kvaliteten i det, de bringer - og dermed sikre en videnskabelig integritet. Dette leder selvfølgelig til en bredere diskussion om, hvem der bærer ansvaret for videnskabelig integritet. Er det forskerne, der laver forskningen? Er det medicinalvirksomheden, der finansierer forskningen? Er det myndighederne, der regulerer forskningen - eller er det de videnskabelige tidskrifter, der formidler forskningen?

Om ikke andet - så vidner første hit fra nedenstående googlesøgning om, at budskabet i blandt andet Noury et al. (2015) trods alt har vundet indpas i Danmark.



5 Diskussion

Vi har indtil nu set tre forskellige cases, der repræsenterede noget vidt forskelligt. I det følgende afsnit vil der med afsæt i de tre cases, diskuteres motiver, hvilke konsekvenser p-hacking har for den videnskabelige integritet, hvis ansvar det er - og hvordan vi mindsker det.

5.1 Motiver og publikationsbias

Forskning skal gerne gøre os klogere på verden og få os tættere på en sandhed, som ellers er svær at bevise. Derfor er det også vigtigt at undersøge, hvorfor nogle forskere ender med at gøre vejen mod sandheden ekstra mudret ved netop at begå p-hacking. Hvilke motiver ligger der bag?

Desværre er der ikke et formelt facit til dette spørgsmål - men ud fra analysen af de tre cases så vi eksempler på både utilsigtet og bevidst p-hacking. Brian Wansink var et eksempel på en forsker, som tydeligvis ikke vidste, hvad han lavede - men som virkelig gerne ville have publiceret sine resultater. Dette står i stærk kontrast til sagen om Study 329, hvor der foregik bevidst manipulation af data, samt en meget selektiv rapportering af resultaterne. Der var en økonomisk interesse i at få studiet publiceret. Samlet set vidner det om i hvert fald tre forskellige motiver.

1. Økonomisk interesse (bevidst)
2. Sjusk og mangel på viden (utilsigtet)
3. Et ønske om at få publiceret resultater

Umiddelbart virker motiv 1 og 2 som rationelle forklaringer på, hvorfor forskere begår p-hacking. De vinder noget på det - eller er ikke klar over det. Til gengæld kan ønsket om at publicere resultater (motiv 3) sandsynligvis være en konsekvens af en større tendens i forskningsmiljøet - nemlig *publikationsbias*. Publikationsbias opstår, når videnskabelige tidskrifter favoriserer resultater, der støtter en alternativ hypotese (altså et statistisk signifikant resultat) - og dermed udelader studier, der ikke kunne vise en effekt. Dette medvirker selvfølgelig til en overrepræsentation af positive resultater, hvilket giver et misvisende billede af den 'sande virkelighed'. Publikationsbias kan i virkeligheden betragtes lidt som tidskrifternes form for p-hacking, da de laver en "selektiv rapportering". Fra tidskrifternes perspektiv, kan der nok være et vis rationale bag, for der er nok ikke ligeså mange læsere, der gider at dykke ned i en artikel, der ikke viser noget. Vi vil ikke læse en artikel om, at fuldmåne ikke viste nogen effekt på motorcykelulykker.

Publikationsbias kan dog potentielt skabe et stærkt motiv for p-hacking, da forskere måske føler et pres for at skulle publicere - med henblik på at fremme deres karriere, opnå anerkendelse eller sikre finansiering til deres næste forskning. Kombinationen af p-hacking og publikationsbias kan derfor i værste tilfælde ende med at forstærke hinanden. Forskere, der er opmærksomme på publikationsbias, kan føle sig presset til at anvende p-hacking - og tidskrifterne kan have tendens til at publicere p-hackede artikler fremfor andre.

Ovenstående motiver er selvfølgelig ren spekulation, da p-hacking er enormt svært at opdage, og dermed er motiverne og tankerne bag p-hacking selvsagt endnu sværere at få indblik i.

5.2 Konsekvenser

I diskussionen af konsekvenserne ved p-hacking kan man se det fra den enkelte borgers perspektiv - men man kan også føre diskussionen op på et højere abstraktionsniveau og diskutere konsekvenserne for den videnskabelige integritet.

I det videnskabelige samfund står vi over for en replikationskrise, hvor en betydelig mængde af tidligere publicerede forskningsresultater har vist sig at være umulige at gentage. Der bliver publiceret alt for mange falsk-positive resultater, som mudrer det samlede billede. Vi så i teoriafsnittet (3.4) en simulation udført af Simonsohn et al. (2011), som viste, at sandsynligheden for at få et falsk-positivt resultat kunne være helt op til 60 %, når man justerede lidt på forskernes frihedsgrader vha. klassiske p-hackingmetoder. Dette betyder altså, at når man begår p-hacking, kan man meget nemt komme til drage de forkerte konklusioner - og dermed have svært ved at gentage studiet.

Når forskere utilsigtet eller bevidst benytter p-hacking for at tilegne sig en p-værdi på under 0.05, bliver objektiviteten og "sandhedsbilledet" udfordret - og p-hacking kan derfor potentielt biddrage til replikationskrisen. Replikationskrisen (med p-hacking som bidragsyder) kan i værste fald skade videnskabens troværdighed - hvilket kan mindske samfundets generelle tillid til forskningsresultater og underminere den rolle, som videnskabelig forskning skal have - nemlig som værende en kilde til *pålidelig* viden. Konsekvenserne af p-hacking kan derfor groft sagt påvirke, hvordan samfundet som helhed opfatter videnskaben.

I de tre forskellige cases var konsekvenserne for de enkelte borgere markant forskellige. I sagen om motorcyklisterne, kan der argumenteres for, at det ikke havde den store konsekvens. Der var måske i værste tilfælde en lidt nervøst anlagt motorcyklist, som lod motorcyklen stå i garagen, når der var fuldmåne - men det er nok også det. I stærk kontrast til dette var der sagen om Study 329. Den havde konsekvenser for mange unge (og deres pårørende), som altså qua konklusionen om den positive effekt af Paroxetin, fik ordineret en medicin, der kunne have ret markante bivirkninger.

Ud fra de ovenstående to tilfælde, kan man derfor godt lave den generalisering, at konsekvenserne af p-hacking kan manifestere sig ved en fejlagtig information, der kan påvirke den enkelte borgers beslutning. Disse beslutninger kan så have en højere eller mindre grad af betydning.

Betydningen af disse beslutninger kan dog være vigtig fra et kollektivt perspektiv. Vi står over for en global klimaudfordring, og har for blot få år siden været vidne til en verdensomspændende covid-19 epidemi. Her er det nødvendigt med beslutninger.

- Hvis vi hver dag skal gå ud med vores grønne bioaffald, spise mindre kød og tage cyklen på arbejde, er det vigtigt, at tilliden til de klimaforskere, der understreger problemets omfang, er på plads.
- Hvis vi kollektivt skal tage imod en ny covid-19 vaccine for at dæmpe epidemiens omfang, er det altafgørende, at vi har en tillid til, at medicinalforskningen ikke har p-hacket resultaterne, da de undersøgte vaccinens effekt og bivirkninger.

Derfor er det oplagt at diskutere, hvordan vi opretholder tilliden til videnskabelig forskning - ved netop at diskustere, hvordan vi så vidt muligt undgår p-hacking.

5.3 P-værdiens magt

I denne opgave har den signifikante p-værdi haft en lidt skjult hovedrolle. Den er prinsessen i Klods Hans, som alle gerne vil have fat i. Brian Wansink testede adskillige hypoteser med flere forskellige responsvariable og fik sin signifikante p-værdi. I sagen om motorcyklisterne, traf forskerne mange (måske heldige) valg, som resulterede i en signifikant p-værdi - og i sagen om Paroxetin var udfaldet det samme. En signifikant p-værdi.

Hvad fortæller det os om p-værdier? Jo, det fortæller os, at p-værdien har utrolig meget magt. Den kan være nøglen til et firmas økonomiske succes, en forskers anseelse og karriere, men nok vigtigst - afgøre om forskningsresultater får bred anerkendelse, jf. tidligere diskussion om publikationsbias. Hvis man ønsker at undgå p-hacking og dermed sikre en vis form for videnskabelig integritet, er det måske derfor oplagt at diskutere, hvorvidt p-værdien fortsat har sin berettigelse og hvordan vi i såfald skal håndtere den.

P-værdi-kultur

I videnskabelige kredse snakker man om en såkaldt p-værdikultur, som er en lidt uhensigtsmæssig kultur, hvor p-værdien anvendes som den primære eller eneste metode til at vurdere statistisk signifikans. Foruden er det forankret i den videnskabelige forskning, at et signifikansniveau på 0.05 symboliserer en skarp skillelinje, hvor man enten kan afvise nulhypotesen eller ej.

I en artikel af Halsey et al. (2015) bliver netop denne problemstilling taget op. De påpeger nemlig, at der er et problem med den måde vi i forskningen bruger p-værdier på. Først og fremmest, at den ret skarpe skillelinje på 0.05 er problematisk, da der statistisk set ikke er den store forskel på en p-værdi på 0.04 og 0.06. Derfor kan man fejlagtigt fortolke resultater som værende betydelige eller ubetydelige - udelukkende baseret på denne (faktisk vilkårlige) grænse.

Halsey et al. (2015) argumenterer derfor for, at denne ret sort-hvide tilgang til p-værdier kan føre til en oversimplificering af data, hvor man måske fejlagtigt kan komme til at overse vigtige nuancer i forskningsresultaterne. Deres hovedpointe er, at p-værdien ikke er så pålidelig, som mange tror - men at den tilgængelig kan variere ret markant. Selvom man laver en undersøgelse korrekt, kan p-værdien være helt anderledes, hvis man prøver at gentage undersøgelsen.

Vi så blandt andet i teoriafsnittet, at både forskellige teststørrelser og forskellige frihedsgrader, kunne betyde en stor forskel i den endelig p-værdi. Desuden så vi i analysen af de tre respektive cases, at p-værdien kunne variere. Dette understreger pointen fra Halsey et al. (2015) om, at man ikke med lukkede øjne skal drage konklusioner alene på baggrund af p-værdien.

"The natural desire for a single categorical yes-or-no decision should give way to a more mature process in which evidence is graded using a variety of measures." (Halsey et al., 2015)

ASA-statement:

For at komme diskussionen omkring p-værdier i møde, valgte American Statistical Association (ASA) at offentligøre seks officielle principper vedrørende brugen af p-værdier (Wasserstein et al., 2016).

Disse principper gik (i en oversat og forkortet version) ud på følgende:

1. **Modelafvigelse:** P-værdier kan vise, hvor uforenelige data er med en specifik statistisk model.
2. **Begrænsninger:** P-værdier måler *ikke* sandsynligheden for, at den undersøgte hypotese er sand eller at data er produceret ved tilfældigheder alene.
3. **Konklusioner:** Videnskabelige konklusioner/beslutninger bør *ikke* udelukkende baseres på, hvorvidt en p-værdi overstiger en bestemt tærskel.
4. **Komplet rapportering:** Alle p-værdier og dertilhørende analyser bør rapporteres - og ikke selektivt udvælges (gennemsigtighed).
5. **Effekt:** En p-værdi måler *ikke* størrelsen af en effekt eller vigtigheden af et resultat.
6. **Evidens:** En p-værdi alene giver *ikke* et godt mål for evidens i forhold til en hypotese. Bredere kontekst og anden evidens er afgørende for at forstå p-værdiens betydning korrekt.

De seks principper er altså en form for brugsanvisning til arbejdet med p-værdier. De forklarer nemlig ret præcist, hvad p-værdien kan - og ligeså vigtigt, hvad den *ikke* kan. Derfor vil det nok ikke være forkert at antage, at ASA's principper også kan være en mulig vej ud af p-hacking-fælden.

Ud fra analysen af de tre cases med p-hacking, så vi nemlig brud på både princip 3, 4 og 6.

ASA understreger i hvert fald vigtigheden i at dele alle resultater. Både de signifikante og ikke-signifikante (princip 4). Dette princip modarbejder direkte p-hacking vedrørende selektiv rapportering - hvor kun resultater, der understøtter en bestemt hypotese, bliver fremhævet. Dette så vi blandt andet i sagen om Paroxetin til unge.

Desuden fremhæver ASA, at statistisk signifikans (fra en p-værdi) ikke nødvendigvis er det samme som videnskabelig evidens (princip 6) - og at man ikke kan drage en konklusion udelukkende baseret på en p-værdi under 0.05. Dette blev fuldstændig ignoreret i sagen om Brian Wansink, da han blev ved med at drage konklusioner ud fra gentagne dataanalyser, hvor p-værdien (som det eneste) netop var under 0.05.

Det står derfor meget klart, at arbejdet med p-værdier kræver en vis nuancering, men at denne nuancering, som ASA beskriver, netop kan være et redskab, som formodentlig kan mindske p-hacking.

Dette leder til næste spørgsmål.

5.4 Hvordan undgår vi p-hacking?

Overskriften er lidt sat på spidsen, for det er nok en illusion at tro, at p-hacking kan elimineres fuldstændigt. P-hacking er som tidligere beskrevet en kompleks størrelse, som i virkeligheden af drevet af menneskelige faktorer - nemlig af de valg, som forskerne træffer i løbet af deres arbejdsproces. Dette vil altid indebære en vis grad af fejlbarlighed, hvorfor fokusset i dette afsnit ikke vil være at undgå p-hacking, men snarere hvordan vi imødekommer og håndterer fænomenet på en konstruktiv måde.

Det blev beskrevet i det tidligere afsnit, hvordan en nuanceret tilgang til arbejdet med p-værdier kan afhjælpe noget af problemet.

En anden metode, som er blevet diskuteret meget er *præregistrering*.

Præregistrering

Præregistrering er en praksis, hvor forskere skal registrere deres plan for studiet *før* de går igang med analysen - herunder deres hypoteser, metoder og strategier. Dette er en god og simpel metode, som vil imødekomme de udfordringer i p-hacking, hvor forskerne træffer valg, som ikke er planlagt - men som blot resulterer i en signifikant p-værdi. Præregistrering kan derfor være en central løsning til at forbedre forskningens pålidelighed og integritet, da det begrænser muligheden for at træffe vilkårlige valg og derfor mindsker sandsynligheden for type-1 fejl (Wicherts et al., 2016). Præregistrering bidrager dermed til en øget gennemsigtighed i forskningsprocessen, som er afgørende for et fænomen som p-hacking, der netop kræver indblik i forskningsprocessen for at blive opdaget.

Det er dog ret essentielt, at præregistreringer bliver indsamlet i en database, som er fuldstændig uafhængig. I sagen om Paroxetin så vi netop et eksempel på, hvorfor vigtige informationer omkring et studie ikke blot skal fremgå i interne dokumenter - men at det skal være en offentlig og uafhængig instans, der fx ikke har interesse i at sælge piller.

Hvis Brian Wansink fra Case I havde præregistreret sine hypoteser vedrørende undersøgelserne fra en pizzarestaurant - ville de endelige fem artikler nok ikke være blevet publiceret. Det står lysende klart i citatet fra hans blogindlæg:

"When she arrived, I gave her a data set of a self-funded, failed study which had null results ... I had three ideas for potential Plan B, C, & D directions (since Plan A had failed)" (Wansink, 2016)

Præregistrering ville fra et generelt perspektiv kunne have en effekt på flere af de metoder vi så i teoriafsnittet (3.4). Vi ville højst sandsynligt se færre tilfælde af 'selektiv rapportering', 'modeltilpasning' og 'manipulation af variabler' - idet forskerne ville gøre det med en vis risiko, da det potentielt kunne blive opdaget ud fra en præregistrering.

I en undersøgelse af Sarafoglou et al. (2022) blev det undersøgt, hvordan prægistrering kan påvirke forskeres arbejdsproces. De adspurgte nemlig 355 forskere (heriblandt forskere med og uden erfaring med præregistrering), hvorvidt de oplevede eller forventede, at præregistrering havde indflydelse på deres forskningsarbejde. Resultaterne viste, at præregistrering generelt blev opfattet positivt - og at størstedelen mente, at det sandsynligvis ville forbedre kvaliteten af deres studier. En væsentlig konklusion var dog også, at præregistrering kan føre til øget arbejdspres og forlængelse af forskningsforløbene - men det var primært de forskere, som ikke havde erfaring med præregistrering.

De beskriver desuden, at præregistrering faktisk også er blevet en forudsætning for at blive publiceret i "New England Journal of Medicine" - som ifølge Sarafoglou et al. (2022) er "verdens mest indflydelsesrige tidsskrift". Dette vidner om et tidskrift, der er bevidst om sit ansvar i at sikre en vis kvalitet i det, de bringer - hvilket faktisk leder til næste afsnit.

Ansvar

Vi så i case III om paroxetin, at tidskriftet JACAAP fortsat ikke havde trukket artiklen tilbage, trods mange anmodninger fra forskellige forskere. Dette fordrede en diskussion om, hvem der bærer ansvaret for at sikre en videnskabelig integritet.

Stephen Hilgartner skelner mellem tre forskellige typer af ansvar, nemlig et politisk, moralsk og kausalt ansvar (Hilgartner, 1990). Kausalt ansvar fokuserer på de underliggende årsager til et fænomen som fx p-hacking. Det kunne være et pres om publicering, mangel på uddannelse, kommerciel interesse osv. Det moralske ansvar placeres derimod hos dem, der kritiseres for problemet (ofte den indiviuelle forsker), hvor politisk ansvar falder på dem, der har magten til at løse problemet. Vi bliver derfor nødt til at forholde os til alle tre typer af ansvar, hvis man vil diskutere, hvordan p-hacking kan mindskes. Forenklet sagt, kan det stilles op således:

- Den individuelle forsker har et moralsk ansvar for at undgå p-hacking.
- Myndigheder, tilsyn og forskningsinstitutter har et politisk ansvar for at undgå p-hacking.
- Identificering af årsager som manglende uddannelse, kommercieller interesser og publikationsbias er et kausalt ansvar for at undgå p-hacking.

Spørgsmålet er så, hvordan vi skal navigere rundt i disse ansvarsområder for effektivt at mindske forekomsten af p-hacking.

For det første bærer den individuelle forsker et moralsk ansvar for at sikre integriteten af sin forskning. I tilfældet med Brian Wansink ser vi konsekvensen af, når forskere (muligvis) lader ønsket om positive resultater eller anerkendelse overskygge en videnskabelig integritet.

For det andet har myndigheder, tilsyn og forskningsinstitutter et politisk ansvar for at skabe fundamentet for en videnskabelig integritet. I Study 329 om Paroxetin blev det klart, at mangel på regulering og overvågning fra myndigheder og institutioner kan medføre, at tvivlsomme forskningsmetoder ikke blot finder sted - men faktisk også bliver offentliggjort.

Med inspiration fra Hilgartner (1990) kan man i arbejdet med at mindske p-hacking regulere på fire forskellige parametre - nemlig straf, kontrol, struktur og uddannelse.

- **Struktur** omfatter klare retningslinjer for datahåndtering og publicering af resultater.
- **Straf** (eller mildere sagt konsekvenserne) for p-hacking kan reguleres. Når forskere ved, at deres overtrædelser vil medføre alvorlige konsekvenser, kan det bidrage til at forskere måske undgår bevidst p-hacking.
- **Uddannelse** er selvfølgelig en oplagt metode, som vi allerede har været lidt inde på. Dette kunne fx. være uddannelse i korrekt brug af statistiske metoder og en nuanceret tilgang til p-værdier (ASA statement Wasserstein et al. (2016)).
- **Kontrol**, som fx. præregistrering og peer review, kan være afgørende for at identificere og rette fejl - inden de bliver godkendt og publiceret.

I sagen om Brian Wansink, kunne man retrospektivt have ønsket at regulere lidt på kontroldelen - for hvordan kan det være, at en så berømt og anerkendt forsker blev opdaget ved en fejl? - og at det faktisk var almindelige mennesker (dog med en statistisk baggrund), der begyndte at grave i hans arbejde?

Disse spørgsmål kunne understrege et større behov for kontrol. Havde der været en strengere kontrol af hans arbejde - herunder præregistrering og peer-review, kunne Wansinks fejlagtige resultater være blevet opdaget og håndteret tidligere. I sagen om Paroxetin kunne man selvsagt også have skruet lidt på struktur- og kontroldelen, da både myndigheder, tidskrifter og læger gik med på konklusionen om, at paroxetin var effektivt mod behandling af unge med depression.
Desuden kunne uddannelse - fx. ved at læse ASA's statement - måske have fået Brian Wansink og forskerne bag motorcyklerne til at stoppe op og tænke: *Kan vi lave denne konklusion alene baseret på en lav p-værdi?* Samlet set vidner det om, at der er mange parametre i spil, når man diskuterer, hvordan man skal mindske p-hacking. Vi kan både betragte p-hacking som et individuelt ansvar - men altså også diskutere, hvorvidt der et strukturelt problem i måden, forskere analyserer data på og hvordan studierne bliver kontrolleret og reguleret.

Vi startede ud med at snakke om, hvilket ansvar tidskrifterne har. Vi så i sagen om paroxetin, at tidskriftet JACAAP endnu ikke har fjernet artiklen af Keller (2001), trods mange anmodninger fra forskere (Newman, 2010). Tidskrifterne er en lidt underlig spiller, for man kan argumentere for at de i virkeligheden repræsenterer lidt af det hele.

- Bidrager til publikationsbias (kausault ansvar)
- Har magten til at gøre noget ved publikationsbias (politisk ansvar)
- Skal sikre, at der er kvalitet i det, de publicerer (moralsk ansvar)

Tidskrifterne har altså et ansvar for at sikre kvaliteten af den forskning, de publicerer. De er også ansvarlige for peer review-processen, der kan fungere som et vigtigt filter inden resultatet når et bredere publikum. På den anden side kan tidskrifter (især dem der drives som private virksomheder) have interesser, der ikke nødvendigvis går hånd i hånd med videnskabelig integritet. De kan være motiverede af profit og opmærksomhed, hvilket måske favoriserer "sensationelle" resultater frem for studier, som ikke viste nogen effekt (publikationsbias).

Samlet set vidner det om, at der er flere knapper at trykke på, hvis man skal sikre en videnskabelig integritet. Der er både et ansvar fra den individuelle forsker for at sikre kvalitet - men myndigheder, tilsyn og tidskrifter har også en en meget stor rolle i at opbygge og vedligeholde et forskningsmiljø, der fremmer god forskningspraksis og etik. Det er dog en balancegang, når man skal navigere i ansvarsområder og eventuel regulering.

Hilgartner (1990) beskriver, hvordan nogle måske kunne frygte, at disse forslag kunne medføre en øget bureaukratisk indgriben i den videnskabelige proces, hvilket kan hæmme kreativiteten og udfordre selvstændigheden. Der er en frygt for, at regulering fra myndighederne kan bane vejen for en større politisk indflydelse på videnskaben.

6 Konklusion

I dette projekt har der været et meget stort forbrug af ord som "sandsynligvis", "muligvis", "måske", "formentlig" og "formodentlig" - og disse blev brugt for netop at manifestere, at der var tale om et lidt usikkert udsagn. Det understreger, hvor subtil og kompleks en størrelse p-hacking er - for vi ved ikke særlig meget om p-hacking. Vi ved dog, at fænomenet eksisterer og at det opstår. Vi kender også til adskillige metoder, der kan medføre p-hacking. Det indbefatter blandt andet 'selektiv rapportering', 'manipulation af variabler', 'fjernelse af outliers', som blev gennemgået i teoriafsnittet - men fælles for dem alle er, at det kræver et indblik i forskernes arbejdsproces, hvis de skal opdages.

I en simulation af Simonsohn et al. (2011) så vi, hvordan man med en vis frihed i forskernes valg, kunne øge sandsynligheden for at få et statistisk signifikant resultat helt op til 60 %, hvilket illustrerer, at omfanget af type-1 fejl godt kan være massivt.

I analysen af de tre forskellige cases, så vi netop eksempler på disse type 1-fejl. Vi fik indsigt i Brians Wansinks arbejdsmetode ud fra et tilfældigt blogindlæg, som afslørede en meget uhensigtsmæssig arbejdsmetode. De manipulerede med responsvariablerne og testede adskillige hypoteser for at opnå en signifikant p-værdi fra et ellers dødsdømt datasæt. Det samme så vi i sagen om Paroxetin til unge, hvor firmaet GSK tilbageholdt vigtige informationer omkring det fejlagtige forsøgsresultat - og alligevel blev der manipuleret med de oprindelige responsvariable, så de fik en statistisk signifikant p-værdi. Desuden blev en selektiv rapportering skyld i, at Paxil blev solgt vidt og bredt i USA.

De tre cases havde det tilfælles, at de cementerede, hvor vigtigt det er at overveje de valg, man træffer i forskningsprocessen - men samtidig så vi også, at både konsekvenserne og motiverne bag var vidt forskellige. De kunne blandt andet være drevet af kommercielle interesser, ønsket om at publicere eller blot have en mangelfuld viden omkring statistisk analyse.

Konsekvenserne varierede også markant - både hvad angår konsekvensen for almindelige borgere, men også konsekvensen for de respektive forskere. Overordnet set kan p-hacking risikere at bidrage til den i forvejen store replikationskrise, som videnskaben står over for - og dette kan potentielt underminere tilliden til videnskabelig forskning.

Der blev introduceret to helt konrekte tiltag (dog med forskellige tilgange) for at imødekomme problemet, nemlig præregistrering og principperne fra "American Statistical Association (ASA)". ASA udgav nemlig seks principper, som skulle være en slags manifest og modsvar til den løbende debat omkring forskeres håndtering af p-værdier - herunder problematikken med den meget skarpe tærskel på 0.05. Dette kunne potentielt forhindre nogle former for p-hacking, da man ud fra principperne ikke må lave en konklusion på baggrund af en p-værdi - uden også at betragte andre faktorer. Foruden anerkendelsen af et strukturelt problem, hvad angår p-værdier, kunne man også i højere grad indføre kravet om præregistrering, hvor forskere inden analysearbejdet skal registrere sin hypotese og fremgangsmåde. Dette er en anden metode, som ville medføre en større gennemsigtighed i arbejdsprocessen - men hvor ansvaret bliver lagt over på den individuelle forsker.

De to metoder har hver sin tilgang - mere uddannelse (ASA) og øget kontrol (præregistrering).

Samlet set, vidner det om, at p-hacking er et begreb, som eksisterer og forekommer. Det er en udfordring, da det medfører en ret stor risiko for type-1 fejl, som mudrer det samlede billede - og

som i værste fald sænker befolkningens tillid til videnskab. Problemet er bare, at det er så svært at opdage. Det kræver derfor en kombineret indsats fra både forskere, forskningsinstitutter, myndigheder og tidsskrifter, hvis man ønsker en gennemsigtig og pålidelig forskning.

Forskere kan tage initiativ til at fremme åbenhed og gennemsigtighed ved fx at præregistrere studier, hvor det er muligt. Forskningsinstitutter og universiteter kan uddanne i god forskningspraksis og etik. Myndigheder og tidsskrifter kan bidrage ved at indføre strengere retningslinjer, hvis man vil publicere - og måske endda indføre regler for præregistreringer.

Om ikke andet har dette projekt forhåbentlig givet en indsigt og en øget bevidsthed om problemet - hvilket er et værktøj i sig selv, hvis man vil undgå et så nyt og uudforsket fænomen som p-hacking.

Litteratur

Anaya, J., Brown, N. J. L. og van der Zee, T. (2017). Statistical heartburn: an attempt to digest four pizza publications from the cornell food and brand lab. *BMC Nutrition*, 3:54.

Bartlett, T. (2017). Spoiled science: How a seemingly innocent blog post led to serious doubts about cornell's famous food laboratory. https://www.chronicle.com/article/spoiled-science/.

Bergström, G., Örjan Ekblom, Frisk, M. K., Fagman, E., Arvidsson, D., Börjesson, M. og Zou, D. (2024). Eveningness is associated with coronary artery calcification in a middle-aged swedish population. *Sleep Medicine*, 113:370-377.

Flyvbjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative Inquiry*, 12, no. 2:219-245.

Halsey, L. G., Curran-Everett, D., Vowler, S. L. og Drummond, G. B. (2015). The fickle p value generates irreproducible results. *Nature methods*, 12:179-185.

Hansen, E. (2012). *Introduktion til Matematisk Statistik*. Institut for Matematiske Fag, Københavns Universitet, 3. udgave.

Hilgartner, S. (1990). Research fraud, misconduct, and the irb. *IRB: Ethics Human Research,*, 12(1):1-4.

Jakobsen, S. E. (2024). Natteravne har næsten dobbelt så stor risiko for åreforkalkning som morgenmennesker. https://videnskab.dk/krop-sundhed/natteravne-har-naesten-dobbelt-saa-stor-risiko-for-aareforkalkning-som-morgenmennesker/.

Jureidini, J. N., McHenry, L. B. og Mansfield, P. R. (2008). Clinical trials and drug promotion: Selective reporting of study 329. *The International journal of risk & safety in medicine*, 20.1-2:73-81.

Keller, M. B. (2001). Efficacy of paroxetine in the treatment of adolescent major depression. *Journal of American Academy of Child and Adolescent Psychiatry (JACAAP)*, 40:7.

Newman, M. (2010). The rules of retraction. *BMJ*, 341:c6985.

Noury, J. L., Jureidini, J., Nardo, J. M., Healy, D., Raven, M., Tufanaru, C. og Abi-Jaoude, E. (2015). Restoring study 329: efficacy and harms of paroxetine and imipramine in treatment of major depression in adolescence. *BMJ (Online)*, 351:h4320.

Redelmeier, D. A. og Shafir, E. (2017). The full moon and motorcycle related mortality: population based double control study. *BMJ*, 359:j5367.

Sarafoglou, A., Kovacs, M., Bakos, B., Wagenmakers, E.-J. og Aczel, B. (2022). A survey on how preregistration affects the research workflow: better science but more work. *R Soc Open Sci*, 9(7):211997.

Simonsohn, U., Nelson, L. D. og Simmons, J. P. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22:1359-66.

Simonsohn, U., Nelson, L. D. og Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of experimental psychology*, 143, no. 2:534-547.

Smith, G. (2023). *Distrust - Big data, data torturing and the assault on science*. Oxford University Press.

Tolver, A. og Hansen, N. R. (2024). *The Mathematics Behind ModernDive*. Institut for Matematiske Fag, Københavns Universitet, 3. udgave.

Unknown (1998). Adolescent depression - position piece on the fase III clinical studies. https://www.industrydocuments.ucsf.edu/drug/docs/#id=xrfw0217. Internt dokument fra GSK.

Wansink, B. (2016). The grad student who never said no. https://archive.ph/cPxmm. Blogindlæg fra arkiv.

Wasserstein, R. L., Schirm, A. L. og Lazar, N. A. (2016). The asa's statement on p-values: Context, process, and purpose. *The american statistician*, 70(2):129-133.

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M. og van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Front. Psychol*, 7:1832.