Economics
Working Papers

2018-7

Frekvensbaserede versus bayesianske metoder i empirisk økonomi

Tom Engsted

Abstract:
Indenfor økonomi og samfundsvidenskab har den klassiske frekvens-baserede analysemetode traditionelt været fremherskende, men de senere år er flere samfundsforskere begyndt at anvende bayesianske metoder i empirisk modellering. I denne artikel beskrives og sammenlignes de to metoder. Der argumenteres for, at vi i højere grad bør anvende den bayesianske tilgang. Den klassiske metode giver sandsynligheden for data, givet modellen (nulhypotesen), mens den bayesianske metode giver sandsynligheden for modellen, givet data. Anvendelse af "p-værdien"i det klassiske hypotesetest fører til for mange "falsk positive" resultater. Den bayesianske metode er mere velegnet end den klassiske til analyse af de hypoteser økonomer arbejder med, hvor en model ikke tilstræbes at være "sand", men i stedet opfattes som en grov approksimation til virkeligheden.

AARHUS BSS
DEPARTMENT OF ECONOMICS
AND BUSINESS ECONOMICS
AARHUS UNIVERSITY

Frekvensbaserede versus bayesianske metoder i empirisk økonomi[1]

Tom Engsted
Institut for Økonomi, Aarhus Universitet
August 2018 (første version marts 2018)

Abstract: Indenfor økonomi og samfundsvidenskab har den klassiske frekvensbaserede analysemetode traditionelt været fremherskende, men de senere år er flere samfundsforskere begyndt at anvende bayesianske metoder i empirisk modellering. I denne artikel beskrives og sammenlignes de to metoder. Der argumenteres for, at vi i højere grad bør anvende den bayesianske tilgang. Den klassiske metode giver sandsynligheden for data, givet modellen (nulhypotesen), mens den bayesianske metode giver sandsynligheden for modellen, givet data. Anvendelse af 'p-værdien' i det klassiske hypotesetest fører til for mange 'falsk positive' resultater. Den bayesianske metode er mere velegnet end den klassiske til analyse af de hypoteser økonomer arbejder med, hvor en model ikke tilstræbes at være 'sand', men i stedet opfattes som en grov approksimation til virkeligheden.

Keywords: Hypotesetest, p-værdi, Bayes-faktor, beslutningsteori, økonomiske modeller.
JEL Codes: C11, C12, C18

1. Indledning.

Den økonomiske videnskab bliver i højere og højere grad empirisk, jf. Angrist, Azoulay, Ellison, Hill and Lu (2017). For 30 år siden publicerede de

[1] Artiklen er baseret på et foredrag om bayesiansk statistik og machine learning, som forfatteren sammen med ph.d.-studerende Nicolaj Mühlbach gav i Det Samfundsøkonomiske Selskab på Aarhus Universitet, november 2017. Tak til Nicolaj for samarbejdet omkring dette foredrag. Også tak til Svend Hylleberg, Viggo Høst, Michael Møller, Carsten Tanggaard og Allan Würtz for kommentarer til en tidligere udgave af artiklen. Ingen af disse personer er ansvarlige for fejl og mangler i artiklen, ligesom de heller ikke nødvendigvis er enige i artiklens budskaber.

økonomiske tidsskrifter jævnligt rene teoretiske bidrag. I dag er det i de brede tidsskrifter sjældent at se en artikel med en teoretisk model, der ikke også udsættes for empirisk validering i artiklen. Den stigende fokus på empiri gør det relevant løbende at diskutere de statistiske metoder, der anvendes i de empiriske analyser. Traditionelt har den klassiske frekvensbaserede tilgang til empirisk analyse været fremherskende, men de senere år har den bayesianske tilgang vundet indpas. Nærværende artikel diskuterer de to tilgange med særlig fokus på anvendelser indenfor samfundsvidenskab.

Blandt statistikere har der siden 1950'erne været en ophedet debat mellem 'bayesianere' og 'frekventister'. Debatten har ikke kun handlet om, hvilken slags statistisk analyse der er mest hensigtsmæssig i konkrete sammenhænge, men også om dybere videnskabssteoretiske og filosofiske problemstillinger, herunder diskussioner om den 'korrekte' opfattelse af sandsynlighesbegrebet.

Som nævnt har den frekvensbaserede tilgang været den fremherskende indenfor de fleste praktiske og empiriske videnskaber - herunder økonomi. Ifl. denne tilgang estimeres og testes en ukendt populationsparameter, θ, på basis af en indsamlet stikprøve. Hypoteser om parameteren formuleres som en nulhypotese, H_o: $\theta = \theta_o$, overfor en alternativhypotese, eksempelvis H_1: $\theta > \theta_o$. En 'teststatistik' konstrueres udfra den estimerede værdi af θ, $\widehat{\theta}$, og udfra teststatistikkens fordeling under H_o, beregnes en såkaldt 'p-værdi'. Hvis denne p-værdi er mindre end et valgt signifikansniveau (typisk 5%), forkastes H_o til fordel for H_1.

Beregning af p-værdien spiller en central rolle i denne analysemetode, men forskere, analytikere og brugere af empiriske undersøgelser er ikke altid klar over, hvad p-værdien egentlig måler. Det er en udbredt misforståelse, at p-værdien udtrykker sandsynligheden for, at H_o er sand. Det er måske ikke så underligt, at denne misforståelse opstår. En lav p-værdi indikerer, at der i statistisk forstand er stærk evidens imod H_o. Det ligger dermed lige for at konkludere, at jo lavere p-værdi, jo mindre sandsynlig er H_o. Men p-værdien er ikke sandsynligheden for at H_o er sand. I den klassiske frekvensbaserede statistik giver det ikke mening at tale om sandsynligheden for H_o (eller H_1). p-værdien udtrykker i stedet sandsynligheden for de observerede stikprøvedata (og dermed test-statistikken), *givet* at H_o er sand!

Hvis man er interesseret i sandsynligheden for H_o, eller de relative odds for H_o versus H_1, er man nødt til at anlægge en bayesiansk tilgang til problemstillingen.

At de klassiske frekvensbaserede tests siger noget om sandsynligheden for at observere data, givet H_o, har betydning for fortolkningen af testene. Ofte

fortolker vi en stor teststørrelse med en lav p-værdi som tegn på, at der er (meget) stærkere evidens for alternativhypotesen H_1 end for H_o. Men en sådan fortolkning er ikke nødvendigvis korrekt. En lav p-værdi udelukker ikke, at der kan være stærkere evidens imod H_1 end imod H_o. Og udfra en bayesiansk synsvinkel kan det sagtens forekomme, at H_o er mere sandsynlig end H_1, selvom H_o forkastes på et 5% signifikansniveau (eller omvendt, at H_o er mindre sandsynlig end H_1, selvom H_o ikke forkastes på 5% niveau). I forlængelse heraf følger det også, at p-værdien ikke kan bruges til at sammenligne modeller. Modellen med den højeste p-værdi er ikke nødvendigvis den, der 'fitter' bedst.

At mange empirikere anvender klassiske hypotesetests og p-værdien uden at kende den præcise fortolkning af disse, har fået statistikere til at tale om "misbruget af p-værdien", et misbrug der lader til at være udbredt i mange videnskaber. Det fik i 2016 *The American Statistical Association* til at udsende et officielt dokument om den korrekte brug af p-værdien: *ASA Statement on Statistical Significance and P-Values* (Wasserstein and Lazar, 2016).

Som nævnt indledningsvis, anvender empiriske økonomer i stigende omfang bayesianske metoder som alternativ (eller supplement) til de klassiske frekvensbaserede metoder. Årsagerne til dette er flere. For det første har fremkomsten af moderne computere gjort det muligt, at foretage de ret tunge beregninger, der ofte (men ikke altid, se nedenfor) kræves i bayesiansk analyse. For det andet, er empirikere i stigende grad blevet opmærksom på, at sandsynligheden for H_o givet stikprøven, $P(H_o \mid \text{data})$, ofte er mere interessant end sandsynligheden for stikprøven givet H_o, $P(\text{data} \mid H_o)$. Og for det tredje - i forlængelse af det foregående punkt - har økonomer efterhånden erkendt, at økonomiske modeller er og altid vil være grove approksimationer til virkeligheden. Det betyder ikke, at modellerne er ubrugelige[2], men det betyder, at test af en 'sand' H_o bliver mindre interessant end at undersøge, hvor godt en - under alle omstændigheder fejlspecificeret - model approksimerer data. Der kan argumenteres for, at den bayesianske tilgang bedre end den klassiske frekvensbaserede tilgang fortæller, hvor godt en fejlspecificeret model fitter data. Jeg vil i denne artikel især fokusere på de sidste to af disse tre punkter.

Jeg har i udarbejdelsen af artiklen ladet mig inspirere af mange ældre og

[2]Som statistikeren George Box udtrykte det i 1978: "*All models are wrong, but some are useful*" (https://en.wikipedia.org/wiki/All_models_are_wrong).

nyere bidrag, men især af Leamer (1978), Startz (2014) og Harvey (2017). Sidstnævnte er præsidenten for *The American Finance Association*'s tale ved foreningens årsmøde i januar 2017 og publiceret i *Journal of Finance*. Harvey's artikel er bemærkelsesværdig derved, at han i denne (og i en tidligere artikel, Harvey, Liu and Zhu, 2015) skriver, at den traditionelle analysemetode har ført til, at "*many of the results being published will fail to hold up in the future*" (Harvey, 2017, p.1399) og "*most claimed research findings in financial economics are likely false*" (Harvey et al., 2015, p.5).[3] Harvey argumenterer for, at vi fremover i højere grad bør inddrage bayesianske metoder i vores empiriske analyser. Som vi skal se, hvis man bruger kriterier, der minder om dem frekventister traditionelt anvender, fører en bayesiansk analyse til færre 'signifikante' resultater, end hvad der følger af den frekvensbaserede analyse.[4]

I afsnit 2 gives en beskrivelse af den klassiske frekvensbaserede tilgang til statistik. Da denne tilgang antages at være velkendt, er beskrivelsen ret kortfattet. Der fokuseres især på p-værdien og det konventionelle valg af et 5% signifikansniveau. I afsnit 3 beskrives, hvordan traditionel beslutningsteori principielt og i praksis kan anvendes til bestemmelse af det 'optimale' signifikansniveau i de klassiske hypotesetests. Dette afsnit fungerer endvidere som en overgang til den bayesianske analyse, der beskrives i afsnit 4. Forskellen mellem objektive og subjektive sandsynligheder diskuteres. Bayes formel beskrives, og det vises, hvordan denne formel kan anvendes til at beregne sandsynligheden for H_o, givet data. Det diskuteres endvidere, hvordan det klassiske test indebærer en implicit prior. Til sidst i dette afsnit beskrives bayesiansk estimation. I afsnit 5 beskrives, hvordan simple bayesianske metoder kan anvendes til at beregne bayesianske p-værdier udfra de klassiske teststatistikker. Metoderne illustreres på konkrete eksempler,

[3] Harvey's budskab blev formuleret mere generelt i en berømt artikel af Ioannidis (2005) med titlen "*Why most published research findings are false*". Indenfor økonomi og økonometri har der naturligvis før Harvey været kastet et kritisk blik på de klassiske signifikanstests, eksempelvis Keuzenkamp and Magnus (1995).

[4] Relaterede emner som ikke diskuteres yderligere i nærværende artikel er "*data mining*" og "*p-hacking*". Andre betegnelser for de samme fænomener er "*data snooping*" og "*data dredging*". Alle empirikere kender til fristelsen til at "torturere dataene til de bekender" så de til slut giver det "ønskede" resultat. Et problem i samme boldgade er tidsskriftredaktørers og referees forkærlighed for "signifikante" resultater. Et andet vigtigt problem, som heller ikke diskuteres yderligere i denne artikel, er den manglende korrektion af signifikansniveauet når der udføres multiple tests; et problem som kendetegner mange empiriske studier.

bl.a. på en politisk meningsmåling og på et eksempel fra den videnskabelige litteratur om forudsigelighed af aktieafkast. Afsnit 6 indeholder nogle metodiske overvejelser om de særlige kendetegn ved økonomiske modeller, og betydningen heraf for empirisk analyse af sådanne modeller. Afsnit 7 indeholder nogle afsluttende kommentarer som konklusion på artiklen.

2. Klassisk frekvensbaseret statistik.

De klassiske metoder daterer sig tilbage til 1920'erne og 30'erne med navne som Ronald Fisher, Jerzy Neyman og Egon Pearson, og tager udgangspunkt i det såkaldte 'likelihood-princip': Givet den statistiske model, er det kun stikprøven der indeholder information om modellens parametre, og likelihoodfunktionen opsummerer al relevant information om stikprøven.[5] De klassiske metoder blev oprindeligt udviklet til analyse af eksperimentelle data og små stikprøver.

I den klassiske tilgang opfattes populationsparametre ikke som stokastiske variable, men derimod som faste - men ukendte - konstanter. Maksimum likelihood estimation af disse parametre fremkommer ved at vælge de parameterværdier, der maksimerer 'sandsynligheden' (*likelihood*) for den givne stikprøve. De estimerede parametre er dermed stokastiske variable, da de er en funktion af de stokastiske stikprøveværdier.

2.1 p-værdien.

Hypotesetest om populationsparametrene sker med anvendelse af de estimerede parametre, men da H_o (og H_1) vedrører populationsparametrene, der ikke er stokastiske, giver det i den klassiske tilgang ikke mening at knytte sandsynligheder til hypoteserne. Testets p-værdi udtrykker ikke sandsynligheden for H_o, men sandsynligheden for at observere stikprøvens testværdi (eller en mere ekstrem værdi af denne), givet at H_o er sand.

Et simpelt eksempel med test på en middelværdi kan illustrere tankegangen: Et test på populationsmiddelværdien μ formuleres som et 'T-test'

[5] I 1960'erne blev det blandt statistikere diskuteret, hvorvidt de klassiske frekvensbaserede tests overhovedet er i overensstemmelse med 'likelihood-princippet' (et emne jeg ikke vil berøre nærmere i denne artikel).

$$T \text{ ratio } = \frac{\overline{Y} - \mu_o}{\sqrt{S^2/n}}, \tag{1}$$

hvor $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ er stikprøveestimatet på μ i en stikprøve af størrelsen n, Y_i'erne er stikprøveværdierne ($i = 1, ..., n$), S^2 er stikprøvevariansen på Y_i, og μ_o er værdien af μ under nulhypotesen, dvs. H_o: $\mu = \mu_o$. Lad alternativhypotesen være H_1: $\mu > \mu_o$.

Hvis Y_i'erne er identisk og uafhængigt normalfordelte, er ovenstående teststatistik t-fordelt under H_o med n-1 frihedsgrader, og p-værdien beregnes som $P(T \text{ ratio} > t \mid H_o)$, hvor t er testværdien i stikprøven. De to ord "under H_o" i ovenstående sætning er vigtige; de betyder, at p-værdien er en betinget sandsynlighed, nemlig sandsynligheden for at observere testværdien (eller en mere ekstrem værdi), givet H_o. Testet giver os altså $P(\text{data} \mid H_o)$, ikke $P(H_o \mid \text{data})$. Hvis man er interesseret i $P(H_o \mid \text{data})$, må man bevæge sig over i den bayesianske statistik (se nedenfor).

Fisher (1925) populariserede anvendelsen af p-værdien som et mål for, hvor stærk evidens der er imod H_o. Men Fisher opererede ikke med en alternativhypotese (H_1) og dermed heller ikke med begreberne Type-1 og Type-2 fejl, eller 'styrken' af et test. '5% reglen' (p-værdi < 0.05 indebærer forkastelse af H_o) stammer oprindeligt fra Fisher, men han pointerede, at reglen ikke skal opfattes som en fast og dogmatisk regel.

Neyman and Pearson (1933) introducerede alternativhypotesen, H_1, eksplicit valg af signifikansniveau, den 'kritiske region', samt Type-1 og Type-2 fejl. Neyman and Pearson opererer derimod ikke med en p-værdi. Man starter med at vælge et signifikansniveau, og dette valg dikterer, om man derefter forkaster eller ikke forkaster H_o til fordel for H_1. For Neyman and Pearson er det konkrete valg mellem H_o og H_1 det centrale, ikke graden af (u)overensstemmelse med H_o målt ved p-værdien, som hos Fisher.

Fisher og Neyman and Pearson var på nogen punkter uenige om den korrekte måde at udføre hypotesetests på (jf. ovenfor), men hos eftertiden er de to tilgange ofte blevet slået sammen til én, og i dag er det ikke ualmindeligt i lærebogsfremstillinger at se principperne bag de klassiske hypotesetests beskrevet som hos Neyman and Pearson, men med beregning af Fishers p-værdi som en integreret del af proceduren.

2.2 Det klassiske 5% signifikansniveau.

Som nævnt stammer det traditionelle valg af et 5% signifikansniveau fra

Fisher, der anbefalede dette niveau som en grov og justerbar regel ved analyse af relativt små stikprøver. Desværre er reglen i eftertiden ofte blevet opfattet og anvendt meget mere rigidt end Fisher anbefalede. Det er vigtigt at pointere, at der intet videnskabeligt eller objektivt belæg er for '5% reglen'. Faktisk afspejler valget af signifikansniveau i høj grad analytikerens subjektive forhåndsvurdering af H_o versus H_1. Valget af et lavt signifikansniveau kan siges at udtrykke analytikerens forkærlighed for H_o. Man taler om, at H_o udtrykker '*the maintained hypothesis*', den hypotese man á priori tror mest på, sådan at der i stikprøven skal være meget stærk evidens imod H_o, før man er klar til at forkaste H_o. Dette kræver et lavt signifikansniveau, svarende til en lav sandsynlighed for fejlagtigt at forkaste en sand H_o (Type-1 fejl). I den forstand er der et subjektivt element indbygget i den klassiske testprocedure i og med at proceduren traditionelt har vægtet minimering af Type-1 fejl højere end minimering af Type-2 fejl (mere om dette nedenfor).[6]

2.3 Problemerne med de klassiske hypotesetests.

Der er (mindst) tre problemer knyttet til ovenstående metode, og især til den måde metoden ofte bliver anvendt på i praksis:

For det første er der mange analyser, hvor det ikke er H_o, men i stedet H_1, der udtrykker analytikerens 'maintained hypothesis'. Eksempelvis økonomen, der har udviklet en teoretisk model, der indebærer, at variablen X påvirker variablen Y. Økonomen, der naturligvis tror på sin model, tester den ved at regressere Y på X, $Y = a + bX + u$, og teste H_o: $b = 0$ overfor H_1: $b \neq 0$. Dvs. et test på at X *ikke* påvirker Y. Økonomen vil her typisk anvende et 5% (eller tilsvarende lavt) signifikansniveau, selvom H_o faktisk ikke udtrykker hans 'maintained hypothesis'.

For det andet vil anvendelsen af et fast signifikansniveau (eksempelvis 5%) altid føre til forkastelse af H_o, når stikprøvestørrelsen bliver tilstrækkelig stor. Dette kan ses i formel (1), hvor T statistikken automatisk stiger, når n øges. Dvs. selv meget små og i økonomisk forstand helt ubetydelige afvigelser fra μ_o bliver statistisk signifikante. Dette har fået statistikere til generelt at anbefale, at lade signifikansniveauet være en aftagende funktion af stikprøvestørrelsen. I dag er det mere reglen end undtagelsen i analyser af eksempelvis mikrodata og højfrekvente finansielle data, at have tusindvis af observationer. Men i mange af disse analyser anvendes fortsat det

[6] Af og til anvendes formuleringen "*maintained hypothesis*" ikke om H_o, men om H_1 (se afsnit 2.3). I dele af den statistiske litteratur bruges betegnelsen også om foreningen (engelsk: "*union*") af H_o og H_1.

traditionelle 5% signifikansniveau. Ofte vil det her være mere relevant at anvende 1% eller 0.5% signifikansniveauer. Omvendt gælder det ved analyser af meget små stikprøver, og med anvendelse af tests med lav styrke (dvs. høj sandsynlighed for en Type-2 fejl), at signifikansniveauer højere end 5% vil være relevante. Men den klassiske statistik giver ingen formel procedure til bestemmelse af et 'passende' signifikansniveau.

For det tredje udtrykker p-værdien i hypotesetestet som nævnt sandsynligheden $P(\text{data} \mid H_o)$. Men for økonomer - og samfundsforskere generelt - vil den omvendt betingede sandsynlighed $P(H_o \mid \text{data})$ ofte være mere interessant. Vores modeller estimeres og testes på ikke-eksperimentelle data, og modellerne er under alle omstændigheder jo blot approksimationer til virkeligheden. Ingen økonomisk model foregiver at at være den 'sande' beskrivelse af virkeligheden. Det kan derfor forekomme bagvendt at tale om sandsynligheden for at observere data, givet at modellen (udtrykt ved H_o) er sand. Det er mere interessant at tale om sandsynligheden for H_o (modellen), givet de observerede data. Men de klassiske hypotesetests giver os ikke denne sandsynlighed. Er man interesseret i $P(H_o \mid \text{data})$, må man anlægge en bayesiansk tilgang til problemstillingen.[7]

Kernen i den begrebsmæssige forskel i de to tilgange kan illustreres ved at betragte test af en populationsparameter, θ, med H_o: $\theta = \theta_o$ og H_1: $\theta \neq \theta_o$. Da H_o og H_1 tilsammen indeholder alle tænkelige værdier af θ, kan man godt umiddelbart få den tanke, at hvis et test giver en lav p-værdi og man derfor vælger at forkaste H_o, må man nødvendigvis mene, at data peger på H_1. Men sådan forholder det sig ikke. Den klassiske analyse giver sandsynlighederne $P(\text{data} \mid H_o)$ og $P(\text{data} \mid H_1)$, men disse to sandsynligheder summerer *ikke* til én. $P(\text{data} \mid H_o)$ kan godt være lav samtidig med at $P(\text{data} \mid H_1)$ er endnu lavere! De omvendt betingede sandsynligheder (de bayesianske), derimod, summerer til én: $P(H_o \mid \text{data}) + P(H_1 \mid \text{data}) = 1$. Dvs. hvis man finder, at $P(H_o \mid \text{data})$ er lav, indebærer det nødvendigvis, at $P(H_1 \mid \text{data})$ er høj, og vice versa. I den forstand tilbyder den bayesianske metode en intuitivt mere naturlig og forståelig procedure til valg mellem H_o og H_1. Dette uddybes i afsnit 4.

[7] At der kan være stor forskel på $P(H_o \mid \text{data})$ og $P(\text{data} \mid H_o)$ kan illustreres ved følgende analogi: Lad H stå for "hængning" og D for "død. Vi kan sikkert hurtigt blive enige om, at sandsynligheden for at dø ved hængning er ret stor, dvs. $P(D \mid H) \approx 100\%$, mens den omvendt betingede sandsynlighed - sandsynligheden for at man er blevet hængt hvis man er død - er ret lille, dvs. $P(H \mid D) \approx 0\%$.

3. Beslutningsteori og valg af signifikansniveau.

Som det er fremgået, findes der ikke et objektivt og videnskabeligt 'korrekt' valg af signifikansniveau, α. Stikprøvestørrelsen, og evt. kendskab til testets styrkeegenskaber, vil ofte kunne guide en i valget af α. I princippet bør valget af α afspejle vægtningen af hhv. Type-1 fejl og Type-2 fejl. I henhold til traditionel beslutningsteori, bør beslutningen om at afvise eller ikke afvise H_o, basere sig på en afvejning af omkostningerne eller konsekvenserne ved at begå hhv. en Type-1 og Type-2 fejl.

Følgende analogi anvendes ofte som illustration: I en retssag er en mand anklaget for mord. Dømmes han skyldig, skal han livsvarigt i fængsel (evt. dødsdømmes). Ifølge almindelige retsprincipper er man uskyldig indtil det modsatte er bevist, og der skal meget stærke beviser - eller indicier - til, for at dømme skyldig. Dette svarer til nulhypotesen H_o: uskyldig, og et meget lavt signifikansniveau, eksempelvis $\alpha = 0.001$, svarende til 0.1% sandsynlighed for at sende en uskyldig livsvarigt i fængsel eller på dødsgangen (Type-1 fejl). Type-2 fejlen svarer her til at frikende en skyldig, hvilket vi generelt opfatter som væsentligt mindre problematisk end at dømme en uskyldig person skyldig.[8]

Leamer (1978, kap. 4) foreslog følgende bayesiansk inspirerede procedure ved valg af signifikansniveau: Minimér (mht. α) det forventede tab ved hypotesetestet, $L = p\alpha L_1 + (1-p)\beta L_2$, hvor $p = P(H_o)$ er á priori sandsynligheden for at H_o er sand, α er sandsynligheden for Type-1 fejl (signifikansniveauet), β er sandsynligheden for Type-2 fejl, og L_1 og L_2 er omkostningerne ved at begå hhv. en Type-1 fejl og en Type-2 fejl.

Problemet i praksis når vi tester statistiske hypoteser er naturligvis at opstille den relevante tabsfunktion: Hvordan vægter vi omkostningerne af hhv. Type-1 og Type-2 fejl? Hvis vi ikke umiddelbart ser os i stand til at foretage en sådan vægtning, kan vi som udgangspunkt antage, at de to typer af fejl vægter lige højt, dvs. $L_1 = L_2$. Hvis vi endvidere ikke har nogen á priori viden om, at H_o er mere eller mindre sandsynlig end H_1, kan vi sætte $p = 0.5$. Hermed bliver tabsfunktionen der skal minimeres: $L = \alpha + \beta$. Anvendes denne procedure i eksempelvis T-testet gennemgået i afsnit 2.1 med H_o: $\mu = 0$, H_1: $\mu = 0.5$, og med en stikprøvevarians på 4, fører det til

[8] Fremkomsten af DNA teknologi har gjort det muligt i visse tilfælde at genundersøge tidligere dødsstraffe. Undersøgelser indikerer, at helt op til 4% af dødsdømte i amerikanske fængsler er uskyldige, jf. Politiken (2014). Dvs. juryer i sager om dødsstraf anvender implicit et 4% signifikansniveau. Et tal jeg personligt finder skræmmende højt!

'optimale' signifikansniveauer på 19% og 11% ved stikprøvestørrelser på hhv. $n = 50$ og $n = 100$, altså noget højere end det konventionelle 5% niveau.[9] Bemærk, at for disse stikprøvestørrelser indebærer det konventionelle valg $\alpha = 0.05$ implicit enten en asymmetrisk vægtning af de to fejltyper og/eller en forhåndsformodning om, at $P(H_o) \neq 0.5$. Det er mit indtryk, at empirikere ofte ikke er opmærksom på disse implicitte implikationer. I en konkret anvendelse af Leamers procedure på tests for enhedsrødder (*unit roots*) i tidsserier, finder Kim and Choi (2017), at det 'optimale' signifikansniveau generelt er noget højere end 5%.

I ovenstående procedure opereres der med sandsynligheden for H_o, $p = P(H_o)$. Som nævnt i afsnit 2 giver en sådan sandsynlighed strengt taget ikke mening i et klassisk frekvensbaseret framework. Derimod giver denne sandsynlighed fint mening i et bayesiansk framework. Lad os derfor nu vende os mod bayesiansk statistik.

4. Bayesiansk statistik.

I bayesiansk statistik opfattes populationsparametre som stokastiske variable. Der er usikkerhed knyttet til disse parametre, og denne usikkerhed beskrives ved sandsynlighedsfordelinger. Grundprincippet i bayesiansk statistik er ved hjælp af Bayes formel at kombinere den information stikprøven giver om givne parametre med en á priori viden (eller holdning), og dermed opnå et udsagn (eller ny holdning), der afspejler en vægtning af á priori informationen med den ny information i stikprøven. I modsætning til den klassiske frekvensbaserede statistik, inddrages der altså her eksplicit en á priori viden, som meget vel kan være subjektivt betinget (heraf ordet "holdning"). Det er naturligvis især dette aspekt, der gør den bayesianske tilgang kontroversiel. I dette afsnit illustreres, hvordan den bayesianske tilgang gør det muligt at beregne sandsynligheden for nulhypotesen, $P(H_o \mid \text{data})$, der - som vi så ovenfor - ofte vil være mere relevant, end p-værdien i det frekvensbaserede framework. Men først redegøres der kort for forskellen i opfattelsen af sandsynlighedsbegrebet mellem frekventister og bayesianere.

[9] Der er her anvendt den sande værdi af variansen S^2 ($\sigma^2 = 4$), hvorved teststørrelsen bliver eksakt standardnormalfordelt, hvilket gør beregning af Type-2 fejlen (β) ligetil. Et 'optimalt' signifikansniveau på 5% fås ved en stikprøvestørrelse på $n = 170$. Med $n = 300$ bliver det 'optimale' signifikansniveau 1.5%.

4.1 Objektiv vs. subjektiv sandsynlighed.

For frekventister er sandsynligheder objektive. Sandsynligheden for en hændelse, A, defineres som

$$P(A) = \lim_{n\to\infty} \left(\frac{m}{n}\right), \tag{2}$$

hvor n er antal gange et eksperiment udføres, og m er antal gange hændelsen A indtræder. Det kunne eksempelvis være et eksperiment, der går ud på at afgøre, om en mønt er ægte. Hændelsen A er at få 'Krone', hvis man kaster mønten én gang. Hvis man kaster mønten mange gange (n stor) og frekvensen $\frac{m}{n}$ går imod $\frac{1}{2}$, kan man konkludere, at mønten er ægte. Sandsynligheden i (2) er 'objektiv' i den forstand, at det kun er eksperimentet (stikprøven), der bestemmer sandsynligheden - der indgår ingen subjektiv á priori vurdering af møntens beskaffenhed ved fastlæggelsen af $P(A)$. Sandsynligheder opfattes som relative frekvenser - heraf betegnelsen 'frekvensbaseret' statistik - og, som det ses, er det en afgørende forudsætning i denne sandsynlighedsopfattelse, at eksperimenter (dvs. stikprøver) er gentagelige (på engelsk: *repeatable*).

Den frekvensbaserede sandsynlighedsopfattelse er naturlig i sammenhænge, hvor eksperimenter (dataudtrækning) gentages mange gange under uændrede vilkår - som eksempelvis ved kast med en mønt eller terning, og i kortspil. Men i andre sammenhænge er den frekvensbaserede sandsynlighedsopfattelse ikke meget bevendt. Hvordan vil en frekventist eksempelvis vurdere sandsynligheden for, at Donald Trump bliver genvalgt som USA's præsident? Hændelsen A lader sig nemt definere, men hvad med m og n? Bayesianere vil ikke have dette problem, for i bayesiansk statistik opfattes sandsynligheder som subjektive, baseret på personlige overvejelser, holdninger og viden. Sandsynlighed er graden af 'tro' (engelsk: *belief*) som et individ knytter til en usikker hændelse, og i bayesiansk statistik udtrykkes denne 'tro' i en á priori sandsynlighedsfordeling, der dernæst kan kombineres med ny indsamlet viden gennem Bayes formel til at få en á posteriori sandsynlighedsfordeling. I eksemplet med Trump, har vi hver især på et givet tidspunkt vores egen á priori opfattelse af Trumps chancer for genvalg, og som tiden går og vi observerer hans gøren og laden, justerer vi denne opfattelse. Matematisk sker denne justering i henhold til Bayes formel, som nu vil blive beskrevet nærmere.[10]

[10]For hardcore frekventister giver det slet ikke mening at tale om den matematiske sandsynlighed for hændelser á la præsident Trumps genvalg. Leamer (1978, kap. 2) diskuterer mere indgående forskellen på objektive og subjektive sandsynligheder, herunder

4.2 Bayes formel anvendt på det klassiske hypotesetest.

Bayes formel angiver den betingede sandsynlighed for hændelsen A, givet hændelsen B, $P(A \mid B)$, ved den omvendt betingede sandsynlighed, $P(B \mid A)$, og forholdet mellem de ubetingede sandsynligheder: $P(A \mid B) = P(B \mid A) \cdot \frac{P(A)}{P(B)}$. Formlen følger af de velkendte regler i basal sandsynlighedsregning, og den accepteres naturligvis som korrekt af både bayesianere og frekventister. Uenighederne omstår i anvendelser af formlen.

Anvendt på det klassiske hypotesetest i afsnit 2, indebærer Bayes formel følgende betingede sandsynlighed for nulhypotesen H_o:

$$P(H_o \mid \text{ data}) \equiv P_{H_o} = P(\text{data } \mid H_o) \cdot \frac{P(H_o)}{P(\text{data})}. \tag{3}$$

I (3) angiver $P(H_o)$ á priori sandsynligheden for H_o, mens $P(\text{data} \mid H_o)$ er den sædvanlige likelihoodfunktion. $P(\text{data})$ er - givet den indsamlede stikprøve - en normaliseringskonstant, der ifl. 'loven om den totale sandsynlighed' kan skrives som $(P(\text{data} \mid H_o)P(H_o)) + (P(\text{data} \mid H_1)P(H_1))$. Denne sandsynlighed kan være besværlig at beregne i praksis (se nedenfor).

På samme vis kan sandsynligheden for alternativhypotesen, $P(H_1 \mid \text{data}) \equiv P_{H_1}$, bestemmes, og denne kan da kombineres med P_{H_o} fra (3) til at få den bayesianske *posterior odds ratio*:

$$\frac{P_{H_o}}{P_{H_1}} = \frac{P(\text{data } \mid H_o)}{P(\text{data } \mid H_1)} \cdot \frac{P(H_o)}{P(H_1)} = (\text{Bayes-faktor}) \cdot (\textit{Prior odds ratio}). \tag{4}$$

Posterior odds ratio i (4) giver den relative sandsynlighed for H_o versus H_1 som produktet af den relative á priori sandsynlighed (*prior odds ratio* $= \frac{P(H_o)}{P(H_1)}$) og Bayes-faktoren, $\frac{P(\text{data} \mid H_o)}{P(\text{data} \mid H_1)}$, der angiver den relative likelihood. Bayes-faktoren fortæller, hvordan stikprøvedata'ene ændrer vores á priori opfattelse af hypoteserne til en ny á posteriori opfattelse. Da det naturligvis gælder at $P(H_o) + P(H_1) = 1$ og $P_{H_o} + P_{H_1} = 1$, fås fra (4):

$$P_{H_o} = \frac{P(\text{data } \mid H_o) \cdot P(H_o)}{(P(\text{data } \mid H_o) \cdot P(H_o)) + (P(\text{data } \mid H_1) \cdot (1 - P(H_o)))}. \tag{5}$$

Hovedproblemet i bayesiansk statistik er valget af prior: Hvordan sættes $P(H_o)$? Det kan ofte være svært i praksis. Hvis man ikke har nogen forhåndsformodning eller viden om, at H_o er mere eller mindre sandsynlig end H_1,

hvorvidt subjektive sandsynligheder opfylder de klassiske sandsynlighedsaxiomer, og om subjektive sandsynligheder overhovedet er 'målelige'.

er standardvalget i bayesiansk analyse: $P(H_o) = P(H_1) = 0.5$, dvs. vi er neutrale overfor de to hypoteser. Hermed forsimples (5) til:

$$P_{H_o} = \frac{P(\text{data} \mid H_o)}{P(\text{data} \mid H_o) + P(\text{data} \mid H_1)}. \tag{6}$$

Som det ses af (4) til (6), indgår betingede sandsynligheder, hvor der betinges på både H_o og H_1, eksplicit i beregningen af P_{H_o} og *posterior odds ratio*. Dette er i modsætning til det klassiske frekvensbaserede approach, hvor p-værdien angiver, hvor stærk evidens der er imod H_o, men hvor der typisk ikke tilsvarende beregnes, hvor stærk evidens der er imod et konkret alternativ. Af og til suppleres p-værdien med beregning af testets styrke for konkrete parameterværdier under H_1, men oftest sker det ikke i frekvensbaserede analyser, og når det sker, er det typisk med udgangspunkt i de konventionelle signifikansniveauer i omegnen af 5%.[11]

4.3 Et simpelt eksempel (fra Startz, 2014).

Antag, at vi ønsker at teste, om en given mønt er ægte, dvs. sandsynligheden for 'Krone' er lig med sandsynligheden for 'Plat'. Dette svarer til at teste, om sandsynlighedsparameteren, θ, i Bernoullifordelingen er lig med 0.5. Dette kan formuleres som et eksakt binominalfordelingstest, eller som et approksimativt normalfordelingstest med en teststatistik á la (1), hvor μ svarer til θ, $\mu_o = \theta_o = 0.5$, og hvor $S^2 = \theta_o(1 - \theta_o) = 0.25$ under H_o. Mønten kastes nu n gange, og maksimum likelihood estimatet på θ fås som $\widehat{\theta} = \frac{m}{n}$, hvor m er antal gange der fås 'Krone'.

I Startz (2014) opereres først med et konkret punkt-alternativ, H_1: $\theta = 0.8$ og en neutral prior, $P(H_o) = P(H_1) = 0.5$ (dvs. ingen forhåndsformodning om, at mønten er ægte eller falsk). Ved anvendelse af (6) fås, at i flere tilfælde, hvor $\widehat{\theta} > 0.5$ og p-værdi < 0.05, er P_{H_o} faktisk større end P_{H_1} (se Figure 4.1 i Startz, 2014). Dvs. selvom det klassiske test forkaster H_o på det konventionelle 5% signifikansniveau, så er - med en neutral prior - H_o mere

[11] Da både $P(\text{data} \mid H_o)$ og $P(\text{data} \mid H_1)$ indgår i (4) til (6), kan formlerne alternativt udtrykkes ved testets *size* og *power*, dvs. ved signifikansniveauet og styrken, se Startz (2014) og Harvey (2017). Disse formler viser det - måske umiddelbart overraskende - fænomen, at jo *lavere* styrke et test har, jo *højere* er sandsynligheden for at H_o er sand hvis man forkaster H_o. Med andre ord: Hvis et test med lav styrke forkaster en hypotese om at en parameter er lig med 0, er der en *høj* sandsynlighed for, at det er en '*false discovery*'! Med á priori even odds for H_o og H_1, indebærer Bayes formel: $P(H_o \mid H_o \text{ forkastes}) = \frac{size}{size+power}$, der er aftagende i *power*.

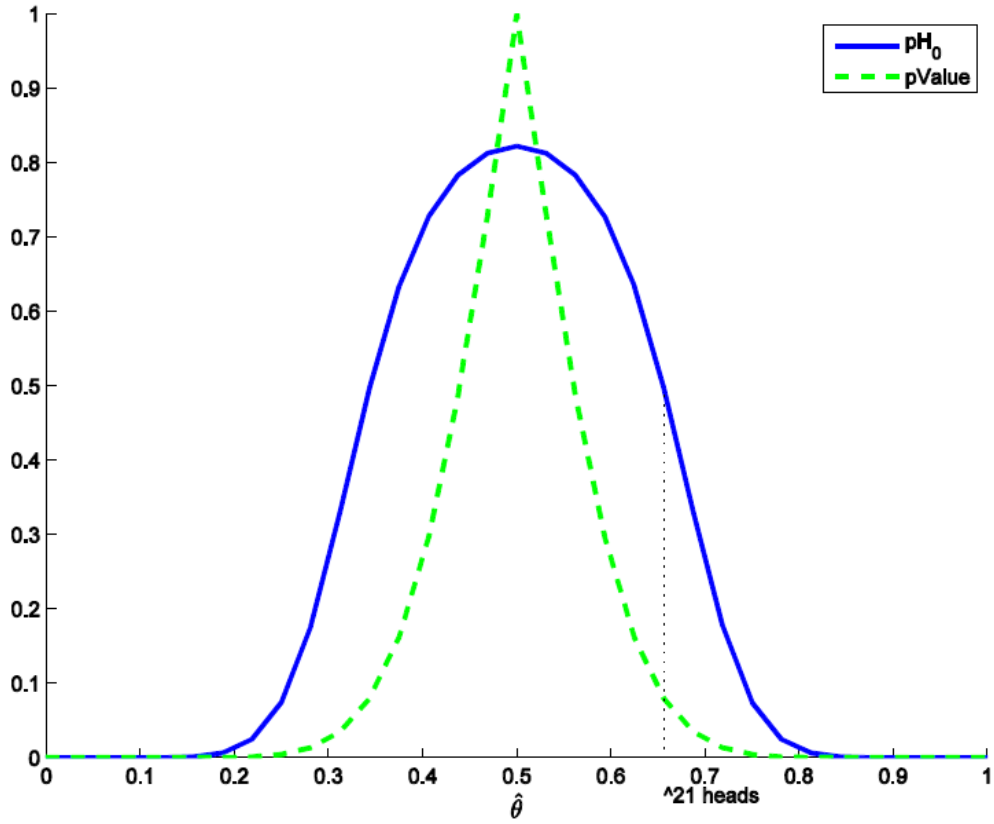


Figure 1: P_{H_o} og p-værdi i møntkasteksemplet. H_o: $\theta = 0.5$, H_1: $\theta \neq 0.5$. $n = 32$. (Fra Startz, 2014, Figure 4.5).

sandsynlig end H_1. Forklaringen er, at selvom der godt nok er stærk evidens imod H_o (lav p-værdi), så er der endnu stærkere evidens imod H_1.

Man kan indvende, at ovenstående eksempel er specielt ved det konkrete punkt-alternativ H_1: $\theta = 0.8$. Det er uden tvivl mere relevant med et traditionelt dobbeltsidet alternativ, H_1: $\theta \neq 0.5$. Men også i dette tilfælde viser Startz, at P_{H_o} ofte vil være væsentligt højere end p-værdien, se Figur 1 (som er Figure 4.5 i Startz, 2014), der for H_o: $\theta = 0.5$ viser henholdvis p-værdien og P_{H_o} for værdier af θ under H_1 over hele enhedsintervallet når mønten kastes $n = 32$ gange. Det ses, at for eksempelvis $m = 21$ 'Krone', svarende til $\widehat{\theta} = 0.656$, fås i det klassiske test en p-værdi på 8%, mens P_{H_o} er 50%.

I øvrigt kompliceres beregningen af P_{H_o} i dette tilfælde væsentligt, da $P(\text{data} \mid H_1)$ i (6) bliver til $\int P(\text{data} \mid \theta_1) f(\theta_1) d\theta_1$, hvor θ_1 er værdien af θ under H_1, og $f(\theta_1)$ er tæthedsfunktionen for θ_1. Dvs. $P(\text{data} \mid H_1)$ er nu det vægtede gennemsnit af likelihood over alle parameterværdier under H_1. Da θ naturligt er begrænset til intervallet mellem 0 og 1, vælger Startz en uniform fordeling i intervallet (0, 1), dvs. $f(\theta_1) = 1$,[12] og han beregner integralet numerisk. Dette illustrerer et generelt kendetegn ved bayesianske metoder, nemlig at de i praksis ofte indebærer numerisk integration, hvilket givetvis er en medvirkende årsag til, at metoderne først de senere år har vundet indpas. Tidligere var beregningerne i bayesiansk analyse simpelthen for krævende, men nutidens moderne computere har elimineret dette problem. I det konkrete tilfælde her behøves faktisk ikke numerisk integration, da der med en uniform prior for θ_1 eksisterer et analytisk udtryk for integralet, se Appendix. (Vi vender tilbage til møntkasteksemplet i afsnit 5.3.1 nedenfor).

4.4 Implicit prior i det klassiske test.

Som nævnt i afsnit 2.2, afspejler valget af et lavt signifikansniveau (typisk $\alpha = 0.05$) i det klassiske test, at analytikeren principielt vægter omkostninger ved Type-1 fejl højere end omkostninger ved Type-2 fejl. Det er derfor ikke korrekt når det af og til hævdes, at den frekvensbaserede tilgang i modsætning til den bayesianske tilgang er 'værdifri' og 'objektiv'. Forskellen i de to tilgange er, at hvor subjektiviteten hos bayesianere gøres eksplicit gennem valget af prior ($p = P(H_o)$ og fordelingen for parameteren under H_1), optræder det subjektive element hos frekventister indirekte gennem valget af signifikansniveau.

Gennem Bayes formel kan frekventistens implicitte prior udledes. Fra det valgte signifikansniveau og teststatistikkens værdi i stikprøven (og dermed p-værdien), kan $P(H_o)$ udledes fra formlerne i afsnit 4.2. Med udgangspunkt i det simple eksempel fra afsnit 4.3, argumenterer Startz (2014) for, at konventionelle beslutningsregler á la "forkast H_o hvis p-værdi $< 5\%$" indebærer implicitte priors, der er alt andet end neutrale og ofte i strid med vores (im-

[12] Læseren undrer sig måske over hvordan P_{H_1} kun bliver 50% når p-værdien for H_o: $\theta = 0.5$ er så lav som 8% og H_1 er $\theta \neq 0.5$. Den lave p-værdi indikerer relativ stærk evidens imod $\theta = 0.5$, og da alternativet indeholder samtlige andre værdier end 0.5, burde P_{H_1} så ikke være væsentlig højere end 50%? Svaret er nej, ikke nødvendigvis, for med en uniform prior for θ_1 er den vægtede likelihood under H_1 også lav. Igen en illustration af betydningen af, hvad der betinges på. Der gælder, at $P(H_o \mid \text{data}) + P(H_1 \mid \text{data}) = 1$, men der gælder naturligvis *ikke*, at $P(\text{data} \mid H_o) + P(\text{data} \mid H_1) = 1$. (Med et andet valg af prior for θ under H_1 kan P_{H_1} blive højere end 50%, se afsnit 5.3.1).

plicitte) forventning. Med Startz's egne ord: "*One of the arguments for using classical testing rather than considering Bayes theorem is that by avoiding specifying a prior we take a neutral approach, a more "scientific" approach, to the data. By considering Bayes theorem we see that all the usual approach does is use an implicit prior, rather than make the prior explicit. We usually think that our standards for significance are chosen precisely to point in the direction of the null unless we have strong evidence to the contrary. But as this example illustrates, our usual standards do not accomplish that goal. In other words, in this example the p-values we usually regard as providing strong evidence against the null and in favor of the alternative do not in fact provide such evidence unless the econometrician already leaned strongly toward the alternative.*" (Startz, 2014, p.141). (Harvey, 2017, afsnit 7, har den samme pointe).

Sammenholdt med diskussionen i afsnit 2.2 ses følgende paradoks: Valget af et lavt signifikansniveau afspejler implicit vores forkærlighed for H_o, så forkastelse af H_o kræver en lav p-værdi. Men en lav p-værdi kan kun tages som udtryk for at sandsynligheden for H_o er lille, hvis vi *på forhånd* har en á priori opfattelse af, at H_o er ret usandsynlig. Hvis læseren er forvirret, er det forståeligt. Paradokset illustrerer, hvor påpasselig man skal være med at kæde p-værdien i det klassiske test sammen med sandsynligheden for H_o.

Tilstræbes der en 'værdifri' ('objektiv') prior, er $P(H_o) = P(H_1) = 0.5$ det naturlige valg. I afsnit 5 vises, hvordan P_{H_o} relativt enkelt kan beregnes for forskellige værdier af $P(H_o)$.

4.5 Bayesiansk estimation.

I den klassiske frekvensbaserede statistik estimeres ukendte parametre ved maksimum likelihood eller tilsvarende metoder, som eksempelvis momentmetoden eller mindste kvadraters metode, og givet modellen (dvs. de valgte variable, funktionsform, fordelingsantagelser, m.v.), er det kun stikprøven, der antages at indeholde information om parametrene. Valget af estimator sker typisk udfra egenskaber som 'forventningsret', 'konsistens' og 'efficiens'. Som det er fremgået ovenfor, fungerer de estimerede parametre som input til hypotesetests om de ukendte populationsparametre.

I den bayesianske tilgang til estimation inddrages der udover stikprøven en á priori opfattelse eller viden om parametrene. Denne á priori opfattelse/viden formuleres i en sandsynlighedsfordeling, der dernæst kombineres med likelihoodfunktionen for stikprøven gennem Bayes formel til at få en á posteriori fordeling for parametrene. Et konkret punktestimat kan da fås

som eksempelvis middelværdien eller medianen i á posteriori fordelingen.

Lad os som illustration vende tilbage til eksemplet i afsnit 4.3. Den ukendte parameter er sandsynlighedsparameteren θ i Bernoullifordelingen. Maksimum likelihood estimatet fås som $\widehat{\theta}_{ML} = \frac{m}{n}$, hvor n er antal gange mønten kastes og m er antal 'Krone' i de n kast. Et Bayes-estimat kan fås ved at antage en á priori fordeling for θ. Da denne parameter naturligt ligger i intervallet 0 til 1, kan en uniform fordeling over dette interval anvendes, $\theta \sim U(0,1)$. Bemærk, at da middelværdi og median i denne fordeling er lig med $\frac{1}{2}$, kan á priori fordelingen siges at afspejle en á priori antagelse om, at mønten er ægte. Antal 'Krone' i de n forsøg følger en binomialfordeling, og kombineres denne med den uniforme á priori fordeling gennem Bayes formel, fås á posteriori fordelingen for θ som en såkaldt *beta-fordeling*. I det konkrete tilfælde her, er middelværdien i beta-fordelingen lig med $\frac{m+1}{n+2}$, der dermed kan anvendes som Bayes-estimator, $\widehat{\theta}_B$ (Barry and Lindgren, 1996, pp. 406-407). Med en kvadratisk tabsfunktion for θ, dvs. en tabsfunktion proportional med $(\theta - \widehat{\theta})^2$, bliver den optimale Bayes-estimator middelværdien i á posteriori fordelingen. Med en tabsfunktion proportional med den absolutte forskel, $\mid \theta - \widehat{\theta} \mid$, bliver den optimale Bayes-estimator medianen i á posteriori fordelingen, jf. Zellner (1971).

Sammenlignes de to estimater, $\widehat{\theta}_{ML} = \frac{m}{n}$ og $\widehat{\theta}_B = \frac{m+1}{n+2}$, ses det, at de er asymptotisk ækvivalente. For $n \to \infty$, er $\widehat{\theta}_B = \widehat{\theta}_{ML}$. Dette er naturligt, da stikprøven for $n \to \infty$ så at sige indeholder uendeligt mere information end á priori informationen. Generelt vil á posteriori fordelingen have normalfordelingen som grænsefordeling når $n \to \infty$, og middelværdien i fordelingen vil være lig med maksimum likelihood estimatet (Zellner, 1971).[13] Omvendt i det modsatte ekstreme tilfælde, hvor der ingen stikprøve foreligger, dvs. $n = 0$. Her er maksimum likelihood estimatoren ikke defineret, mens Bayes-estimatoren blot er givet ved middelværdien i á priori fordelingen, $\frac{1}{2}$.

Kastes mønten én gang, $n = 1$, og resultatet bliver 'Krone' ($m = 1$), fås estimaterne $\widehat{\theta}_{ML} = 1$ og $\widehat{\theta}_B = \frac{2}{3}$. Bliver det derimod 'Plat', fås $\widehat{\theta}_{ML} = 0$ og $\widehat{\theta}_B = \frac{1}{3}$. Intuitionen er ligetil: Med kun én stikprøveobservation (ét møntkast), må maksimum likelihood estimatet nødvendigvis blive enten 1 eller 0. Bayes-estimatet, derimod, bliver op- eller nedjusteret ift. á priori opfattelsen $\theta = \frac{1}{2}$, afhængig af, om det bliver 'Krone' eller 'Plat'. For $n = 1$,

[13] Dette er i modsætning til hypotesetests, hvor den bayesianske p-værdi, P_{H_o}, generelt *ikke* vil gå imod den klassiske p-værdi for $n \to \infty$. (Se afsnit 5).

vægter á priori opfattelsen højt ift. stikprøven. Men, som det ses, jo større n bliver, jo mere vægter stikprøven i forhold til á priori opfattelsen, og jo tættere kommer Bayes-estimatet på maksimum likelihood estimatet.

5. En kombination af den frekvensbaserede og den bayesianske tilgang.

Hovedbudskabet fra bayesianere er, at sandsynligheder nødvendigvis må opfattes som subjektive, og at sandsynligheden for en hændelse A, givet informationen i data, $P(A \mid \text{data})$, afhænger af vores á priori opfattelse af sandsynligheden, $P(A)$. Jo højere (lavere) sandsynlighed vi på forhånd tildeler en hændelse, desto højere (lavere) sandsynlighed har hændelsen, alt andet lige. Dette følger direkte fra Bayes formel: $P(A \mid \text{data}) = P(\text{data} \mid A) = \frac{P(A)}{P(\text{data})}$.

Frekventister accepterer naturligvis Bayes formel som korrekt, men har alligevel ofte svært ved at acceptere ovenstående budskab. En af bayesianismens pionerer, Leonard Savage, gav i 1962 følgende eksempel for at overbevise en skeptiker (jf. Greenhouse, 2012). Betragt følgende tre situationer: 1) En ældre dame påstår, at hun kan skelne mellem, om mælken hældes i tekoppen før eller efter det varme tevand hældes i. 2) En musikprofessor med ekspertise i 1700-tallets komponister påstår, at han kan skelne mellem noder skrevet af Mozart og Haydn. 3) En beruset bargæst påstår, at han kan forudsige udfaldet af kast med en mønt. I alle tre eksperimenter rammer personerne rigtigt i 10 ud af 10 tilfælde. p-værdien på 0.001 ($=0.5^{10}$) er den samme i alle tre tilfælde og fører efter almindelig praksis til forkastelse af nulhypotesen om ingen evne til at skelne. Men de fleste af os kan givetvis blive enige om, at hvor forkastelsen i tilfælde 2) blot bekræfter os i vores forhåndsformodning om, at en musikprofessor naturligvis kan skelne Mozart fra Haydn, gør forkastelsen i tilfælde 3) os meget skeptiske. Vi er selv med en p-værdi på 0.001 ikke overbeviste om, at bargæsten er clairvoyant. Vi vil nok som minimum kræve en gentagelse af eksperimentet før vi lader os overbevise. I tilfælde 1) vil nogen givetvis blive overbevist af forsøget, mens andre ikke vil. Som det ses, afhænger sandsynligheden (rettere: vores opfattelse af sandsynligheden) for en hændelse af vores á priori opfattelse af sandsynligheden.

Hvis man accepterer ovenstående argumentation, og derudover accepterer, at $P_{H_o} = P(H_o \mid \text{data})$ er (mindst) lige så relevant som $P(\text{data} \mid H_o)$, bør man fange interesse for den bayesianske tilgang til statistisk analyse.

En full-blown bayesiansk analyse kræver detaljeret specifikation af á pri-

ori fordelingen for de forskellige hypoteser, efterfulgt af beregning af á posteriori fordelingen, hvilket kan være ret udfordrende i praksis. Og som vi så i afsnit 4.3, selv hvis vi præsætter á priori sandsynlighederne $P(H_o)$ og $P(H_1)$, kræves der stadig generelt specifikation af en á priori fordeling for parameteren under H_1. Statistikere har derfor udviklet simple bayesianske metoder til at beregne P_{H_o} baseret på de velkendte tests fra den frekvensbaserede analyse, og hvor der ikke kræves eksplicit specifikation af á priori fordelingen. Startz (2014) og Harvey (2017) anbefaler, at man som minimum supplerer sine analyser med disse P_{H_o} værdier. I det følgende redegøres for disse metoder.

Husk fra ligning (4), at Bayes-faktoren angiver den relative likelihood for data, givet H_o, versus likelihood for data, givet H_1: $\text{BF} = \frac{P(\text{data}\mid H_o)}{P(\text{data}\mid H_1)}$. Fra (4) fås dermed P_{H_o}, som Harvey (2017) benævner "*Bayesianized p-value*":

$$P_{H_o} = \text{BF} \cdot \frac{P(H_o)}{1 + [P(H_o) \cdot (\text{BF} - 1)]} \tag{7}$$

For at beregne P_{H_o} skal man vælge en Bayes-faktor (BF) og angive en á priori sandsynlighed for H_o ($P(H_o)$). Nedenfor skitseres tre forskellige valg af BF (BIC-BF, MBF og SD-MBF). Vedrørende á priori sandsynligheden for H_o, er $P(H_o) = P(H_1) = 0.5$ det neutrale valg. Man kan sige, at dette valg modsvarer frekventistens ideal om kun at 'lade data tale', uden at pålægge analysen en subjektiv forhåndsformodning. I forlængelse heraf kan man sige, at frekventistens valg af et 5% signifikansniveau i den bayesianske analyse svarer til, at H_o kun forkastes, hvis á posteriori sandsynligheden, P_{H_o}, er mindre end 5%. Dette diskuteres yderligere nedenfor.

5.1 Bayesian Information Criterion (BIC-BF).

Økonometrikere og de fleste empirikere er bekendt med BIC, der også kaldes for *Schwarz Information Criterion*. Kriteriet anvendes - også af frekventister - i mange sammenhænge til valg af modelspecifikation, eksempelvis lag-længde i VAR modeller. Det er mindre velkendt, at kriteriet kan anvendes til at beregne P_{H_o} udfra de klassiske hypotesetests. I modeller, hvor hypotesen kan testes med det sædvanlige T-test, kan BIC og tilhørende værdi af Bayes-faktoren skrives som (jf. Kass and Raftery, 1995; Startz, 2014):

$$\text{BIC} = \ln(n) - t^2, \qquad \text{BIC-BF} = \exp(0.5 \cdot \text{BIC}), \tag{8}$$

hvor t er t-statistikken for H_o og n som tidligere er antal observationer i stikprøven.[14]

Den praktiske fordel ved BIC-BF er, at der ikke kræves en eksplicit specifikation a á priori fordelingen under H_1. Der ligger naturligvis en implicit prior gemt i BIC, og det har vist sig at BIC er ret konservativ i den forstand at der kræves stærk evidens i stikprøven imod H_o for at P_{H_o} bliver væsentlig mindre end P_{H_1}. Som følge heraf skriver Raftery (1999), at i modeller, hvor H_o er, at en parameter er lig med 0 (dvs. ingen effekt), hvis BIC finder evidens for en effekt, så kan vi være ret sikre på, at der *er* en effekt. Men han anbefaler også, at BIC anvendes som en indledende "*baseline reference analysis*" og ikke som det endegyldige valg mellem H_o og H_1.

5.2 Minimum Bayes Factor (MBF og SD-MBF).

Harvey (2017) anbefaler brugen af den såkaldte '*Minimum Bayes Factor*' (MBF). MBF er den nedre grænse blandt alle Bayes-faktorer og er dermed den Bayes-faktor, der giver den stærkeste evidens imod H_o.[15] MBF er i den forstand det diamentrale modstykke til BIC-BF, der - som gennemgået ovenfor - er ret konservativ og derfor vil tendere til at favorisere H_o. Lad som før t være t-statistikken for hypotesen; *Minimum Bayes Factor* beregnes da som MBF $= \exp(-t^2/2)$, der kan indsættes for BF i formel (7).

Harvey (2017) anbefaler også brugen af hvad han kalder SD-MBF, der står for '*Symmetrically and Descending Minimum Bayes Factor*'. SD-MBF er den nedre grænse blandt alle Bayes-faktorer der pålægger en á priori fordeling under H_1, der er symmetrisk og aftagende omkring H_o. Den beregnes som SD-MBF $= -2.718 \cdot$ p-værdi $\cdot \ln($p-værdi$)$, hvor p-værdien som sædvanlig beregnes udfra T-testet under H_o.[16] Da MBF er den nedre grænse blandt

[14] En underliggende antagelse er, at stikprøven er tilstrækkelig stor til at t-statistikken kan approksimeres ved standard-normalfordelingen. Samme antagelse gøres ved beregning af P_{H_o} udfra MBF og SD-MBF, jf. afsnit 5.2.

[15] MBF stammer fra Edwards, Lindman and Savage (1963). MBF indebærer, at á priori fordelingen under H_1 er koncentreret i maksimum likelihood estimatet. Hermed minder MBF meget om det klassiske *likelihood ratio* (LR) test. Forskellen er, at LR testets fordeling er udledt under H_o og giver dermed $P(\text{data} \mid H_o)$. For Bayes-faktorer generelt gælder det, at likelihood under H_o sættes i forhold til et vægtet gennemsnit af likelihood over alle parameterværdier under H_1. En yderligere forskel er, at hvor LR testet automatisk indebærer, at modellen i H_o er nestet under modellen i H_1, kan Bayes-faktorer sammenligne både nestede og ikke-nestede modeller.

[16] Anvendelsen af SD-MBF forudsætter, at p-værdien er mindre end $e^{-1} = 0.368$, et forbehold der ikke nævnes i Harvey (2017), jf. Bayarri and Berger (1998) og Sellke, Bayarri and Berger (2001), der er ophavsmændene til SD-MBF.

alle tænkelige Bayes-faktorer, vil P_{H_o} beregnet under SD-MBF være højere end P_{H_o} beregnet under MBF.

Antag eksempelvis, at et test giver en test-statistik på 1.96 med en p-værdi på 5%, dvs. H_o afvises lige akkurat på det sædvanlige 5% signifikansniveau. Dermed fås MBF $= \exp(-1.96^2/2) = 0.146$. Med en neutral á priori sandsynlighed på $P(H_o) = P(H_1) = 0.5$, beregnes á posteriori sandsynligheden for H_o ifl. (7) til $P_{H_o} = 0.146 \cdot \frac{0.5}{1+[0.5(0.146-1)]} = 0.127$, dvs. 12.7%. Anvendes SD-MBF $= -2.718 \cdot 0.05 \cdot \ln(0.05) = 0.407$, fås $P_{H_o} = 0.289$, dvs. 28.9%. Bemærk hvordan kombinationen af en á priori opfattelse og informationen i stikprøven fører til en revidering af den relative sandsynlighed for de to hypoteser: Vi starter med en neutral prior, $P(H_o) = 0.5$. Bayes-faktoren angiver den relative likelihood for H_o vs. H_1 i stikprøven (0.146 for MBF og 0.407 for SD-MBF). Det ses, at informationen i data peger på H_1 fremfor H_o, så stikprøven fører til en nedjustering af vores opfattelse af sandsynligheden for H_o (P_{H_o}), fra 50% til hhv. 12.7% og 28.9% for MBF og SD-MBF.

Det fremgår, at P_{H_o} i begge tilfælde er højere end p-værdien, men mindre end 50%. Dvs. den bayesianske analyse indikerer, at H_o er mindre sandsynlig end H_1, men evidensen imod H_o vil nok for de fleste ikke forekomme så stærk, som p-værdien på 5% indikerer. Hvis man - som i den traditionelle tankegang - har H_o som '*maintained hypothesis*', og derfor ønsker meget stærk evidens i stikprøven imod H_o for at vælge H_1 (jf. Neyman-Pearson tilgangen som beskrevet i afsnit 2.1), vil P_{H_o} værdier på 12.7% og 28.9% hos mange givetvis ikke føre til forkastelse af H_o. Ikke mindst i lyset af, at MBF og SD-MBF i forvejen har indbygget á priori fordelinger under H_1, der tenderer til at favorisere H_1. Som nævnt anbefaler Harvey (2017) brugen af både MBF og SD-MBF, men ifl. Berger and Sellke (1987), Berger and Delampady (1987) og Kass and Raftery (1995) vil MBF ofte give en for stærk favorisering af H_1. Mindre aggressive Bayes-faktorer (som eksempelvis SD-MBF) vil være at foretrække.[17]

5.3 Illustrationer af de simple bayesianske metoder.

I dette afsnit illustreres de ovenfor beskrevne metoder på det tidligere beskrevne møntkasteksempel, et eksempel med en politisk meningsmåling, og sluttelig et eksempel fra den videnskabelige litteraur om afkastforudsigelighed i ak-

[17] Sammenlignes MBF $= \exp(-t^2/2)$ med BIC-BF $= \exp((\ln(n) - t^2)/2)$, ses det, at de er sammenfaldende for $n = 1$. En illustration af MBF's aggressivitet. For $t = 1.96$ (p-værdi = 0.05) er SD-MBF sammenfaldende med BIC-BF for $n \approx 8$.

tiemarkedet. Sidstnævnte eksempel illustrerer betydningen af á priori holdningen til den testede hypotese for konklusionen på testet.

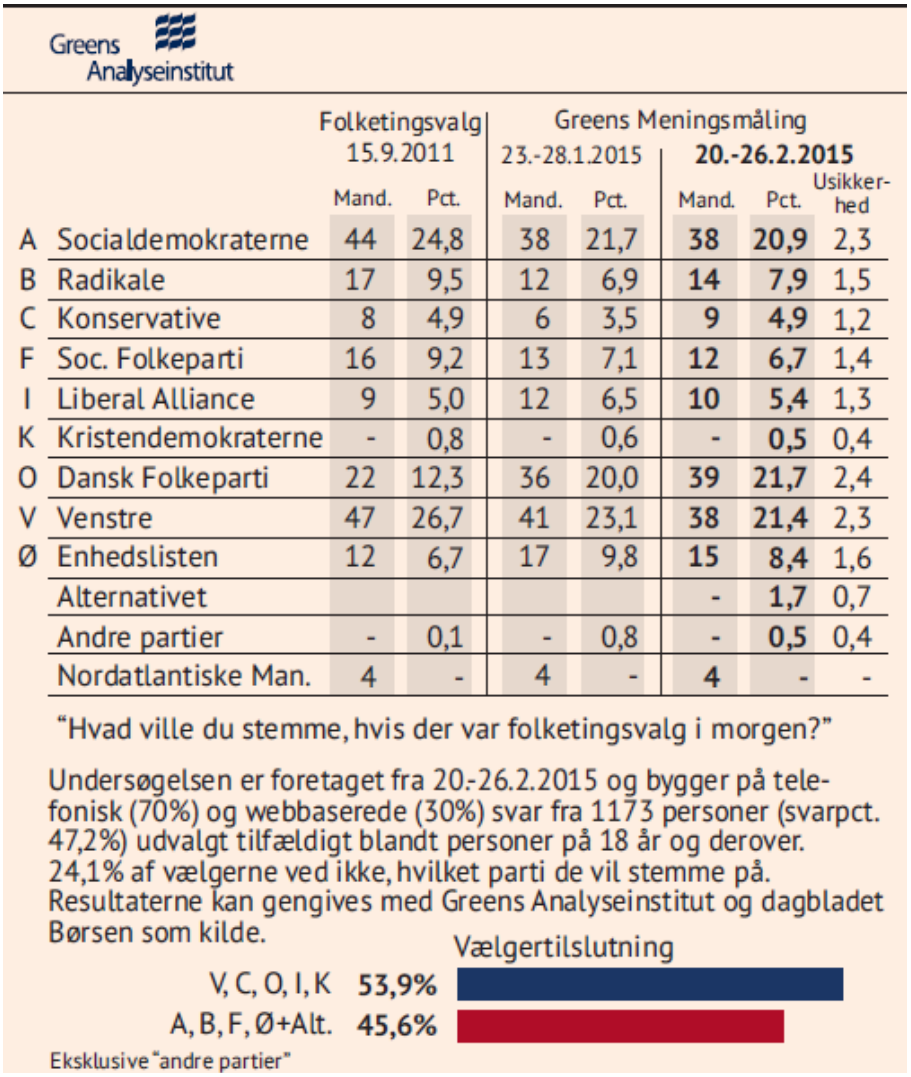
5.3.1 Møntkasteksemplet.
Lad os starte med at anvende metoderne på eksemplet fra afsnit 4.3, hvor en mønt kastes 32 gange og hypotesen om en ægte mønt ($\theta = 0.5$) testes overfor alternativet $\theta \neq 0.5$. Estimatet på sandsynlighedsparameteren er $\widehat{\theta} = 0.656$, test-statistikken er 1.767 og p-værdien er 0.08. Med en uniform prior i intervallet (0, 1) for θ under H_1, er $P_{H_o} = 0.50$ (jf. afsnit 4.3 og Appendix).

Der kan argumenteres for, at en uniform prior for θ er urealistisk da det indebærer, at værdier af θ tæt på 0 eller 1 er lige så sandsynlige som værdier tæt på 0.5. Hermed fås en lav værdi af den vægtede likelihood under H_1, og dermed en relativ lav P_{H_1} værdi. Men anvendes i stedet BIC kriteriet, fås med anvendelse af (7) og (8): BIC $= \ln(32) - 1.767^2 = 0.343$, BIC-BF $= \exp(0.5 \cdot 0.343) = 1.187$ og $P_{H_o} = 1.187 \cdot \frac{0.5}{1+[0.5\cdot(1.187-1)]} = 0.543$, dvs. en endnu højere sandsynlighed for H_o end med den uniforme prior. Eksemplet illustrerer BIC's konservative natur.

Ønskes i stedet en prior, der giver stærkest mulig evidens imod H_o, kan MBF anvendes. Her fås MBF $= \exp(-1.767^2/2) = 0.210$ og $P_{H_o} = 0.210 \cdot \frac{0.5}{1+[0.5(0.210-1)]} = 0.173$ (for $P(H_o) = 0.5$). Som det ses, fås nu stærkere evidens imod H_o, men P_{H_o} er stadig væsentligt højere end den klassiske p-værdi på 0.08. Anvendes SD-MBF, der også indebærer relativ stærk á priori evidens imod H_o, men ikke så stærk som MBF, fås SD-MBF $= -2.718 \cdot 0.08 \cdot \ln(0.08) = 0.549$ og $P_{H_o} = 0.351$.

Om disse á posteriori sandsynligheder er tilstrækkeligt lave til, at man forkaster H_o, er et subjektivt valg hver enkelt må gøre sig. Men bemærk, at dette valg ikke er mere subjektivt, end det valg frekventisten må gøre på baggrund af den klassiske p-værdi på 8%. Bayesianeren vil hævde, at P_{H_o} værdien er et mere naturligt - og forståeligt - mål for den relative evidens for H_o vs. H_1, mens frekventisten vil afvise, at det giver mening at tale om 'sandsynligheden for H_o' og derfor vil holde fast i p-værdien. For en pragmatiker vil det være naturligt at rapportere både p-værdien og P_{H_o} værdierne, og så lade det være op til læseren selv at vurdere evidensen.

5.3.2 Politisk meningsmåling.
Nyhedsmedierne bringer jævnligt politiske meningsmålinger udført af diverse analyseinstitutter, og der har de senere år været noget diskussion om den



Greens Analyseinstitut

		Folketingsvalg 15.9.2011		Greens Meningsmåling 23.-28.1.2015		**20.-26.2.2015**		
		Mand.	Pct.	Mand.	Pct.	Mand.	Pct.	Usikker-hed
A	Socialdemokraterne	44	24,8	38	21,7	**38**	**20,9**	2,3
B	Radikale	17	9,5	12	6,9	**14**	**7,9**	1,5
C	Konservative	8	4,9	6	3,5	**9**	**4,9**	1,2
F	Soc. Folkeparti	16	9,2	13	7,1	**12**	**6,7**	1,4
I	Liberal Alliance	9	5,0	12	6,5	**10**	**5,4**	1,3
K	Kristendemokraterne	-	0,8	-	0,6	-	**0,5**	0,4
O	Dansk Folkeparti	22	12,3	36	20,0	**39**	**21,7**	2,4
V	Venstre	47	26,7	41	23,1	**38**	**21,4**	2,3
Ø	Enhedslisten	12	6,7	17	9,8	**15**	**8,4**	1,6
	Alternativet					-	**1,7**	0,7
	Andre partier	-	0,1	-	0,8	-	**0,5**	0,4
	Nordatlantiske Man.	4	-	4	-	**4**	-	-

"Hvad ville du stemme, hvis der var folketingsvalg i morgen?"

Undersøgelsen er foretaget fra 20.-26.2.2015 og bygger på telefonisk (70%) og webbaserede (30%) svar fra 1173 personer (svarpct. 47,2%) udvalgt tilfældigt blandt personer på 18 år og derover. 24,1% af vælgerne ved ikke, hvilket parti de vil stemme på. Resultaterne kan gengives med Greens Analyseinstitut og dagbladet Børsen som kilde.

Vælgertilslutning

V, C, O, I, K **53,9%**

A, B, F, Ø+Alt. **45,6%**

Eksklusive "andre partier"

Figure 2: Politisk meningsmåling, Børsen 2. marts 2015.

statistiske usikkerhed i disse målinger, ikke mindst fordi målingerne i flere tilfælde har ramt skævt ift. folketingsvalgene.

Figur 2 viser en tabel med en måling fra Greens Analyseinstitut, bragt i dagbladet Børsen den 2 marts 2015. I noten til tabellen rapporteres at stikprøvestørrelsen er 1173 personer, og sidste kolonne viser "Usikkerhed" for hvert politiske parti. Disse usikkerheder er beregnet som klassiske 95% konfidensintervaller. Eksempelvis for Socialdemokraterne (A):

$$1.96 \cdot \sqrt{\frac{0.209 \cdot (1 - 0.209)}{1173}} = 0.023 \quad (2.3 \text{ procent})$$

I mediernes dækning af målingerne bliver den statistiske usikkerhed ofte ikke inddraget når der kommenteres på, om et parti er gået frem eller tilbage ift. seneste måling eller folketingsvalg. Men af og til ser man henvisninger til denne usikkerhed, og man kan faktisk også i aviserne støde på formuleringer á la "fremgangen fra seneste måling er statistisk signifikant".

Hvordan ser en frekvensbaseret analyse og en bayesiansk analyse af en sådan meningsmåling ud? I Tabel 1 har jeg for alle partier beregnet det klassiske Z test (approksimative normalfordelingstest), og tilhørende p-værdi (beregnet for det dobbeltsidede alternativ), for hypotesen, at andelen, der vil stemme på det pågældende parti, er lig med andelen ved folketingsvalget den 15 september 2011. Der tages udgangspunkt i målingen (stikprøven) fra februar 2015. Det ses, at på det konventionelle 5% signifikansniveau, er der statistisk belæg for at sige, at partierne A, F og V er gået tilbage, mens partierne O og Ø er gået frem. For de resterende partier (B, C, I og K) er ændringen ikke statistisk signifikant.

Tabellen indeholder også bayesianske p-værdier beregnet udfra BIC-BF, MBF og SD-MBF, og med neutrale á priori sandsynligheder ($P(H_o) = P(H_1) = 0.5$). For mange af partierne er konklusionen baseret på de bayesianske p-værdier i store træk identisk med konklusionen baseret på de klassiske p-værdier. Men for enkelte partier er der afvigelser. For Enhedslisten (Ø) er den klassiske p-værdi 2%, mens ingen af de bayesianske p-værdier er under 5%. For dette parti vil konklusionen baseret på det klassiske test være, at partiet er gået frem, mens konklusionen baseret på den bayesianske analyse er mindre entydig.

	Klassisk test		BIC-BF		MBF		SD-MBF	
	Z test	p-værdi	BF	P_{H_o}	BF	P_{H_o}	BF	P_{H_o}
A	-3.09	0.002	0.287	0.223	0.008	0.008	0.034	0.033
B	-1.87	0.062	5.972	0.857	0.174	0.148	0.469	0.319
C	0.00	1.000	34.25	0.972	1.000	0.500	-	-
F	-2.96	0.004	0.426	0.299	0.012	0.012	0.060	0.057
I	0.63	0.530	28.10	0.966	0.821	0.451	-	-
K	-1.15	0.248	17.62	0.946	0.514	0.340	0.940	0.485
O	9.80	0.000	0.000	0.000	0.000	0.000	0.000	0.000
V	-4.10	0.000	0.008	0.008	0.000	0.000	0.000	0.000
Ø	2.33	0.020	2.274	0.695	0.066	0.062	0.213	0.175

Note: Á posteriori sandsynligheden P_{H_o} (den bayesianske p-værdi) er beregnet som i formel (7) med $P(H_o) = 0.5$. For C og I er SD-MBF ikke defineret da p-værdien > 0.368 (jf. fodnote 16).

Tabel 1: Klassiske og bayesianske tests af H_o: Ingen ændring ift. folketingsvalget i 2011.

5.3.3 Er aktieafkast forudsigelige?

Et hot emne i den empiriske finansieringslitteratur er, hvorvidt aktieafkast indeholder forudsigelige elementer. Cochrane (2011) giver et survey over denne litteratur og han præsenterer empiriske resultater for USA, hvor afkast regresseres på dividende-pris ratioen. Det er velkendt fra finansieringslitteraturen, at hvis aktieafkast er (delvist) forudsigelige, bør denne ratio kunne fange denne forudsigelighed (jf. Campbell and Shiller, 1988).

På årlige data for perioden 1947-2010 regresserer Cochrane ét-års log afkast, r_t, på den laggede log dividende-pris ratio, dp_{t-1}, og får følgende resultat (Cochrane, 2011, Table III):

$$r_t = \text{ konstant } + 0.13 \cdot dp_{t-1} \tag{9}$$

$$R^2 = 0.10, \quad t\text{-statistik} = 2.61, \quad \text{p-værdi} = 0.009$$

Det ses, at regressionskoefficienten på dp_{t-1} har en t-statistik på 2.61 med en p-værdi på under 1%, dvs. hypotesen om ingen afkastforudsigelighed forkastes klart med det klassiske test. Cochranes analyse er rent frekvensbaseret; han rapporterer ingen bayesianske P_{H_o} værdier. Jeg har derfor i

Tabel 2 beregnet P_{H_o} værdier udfra ovenstående t-statistik og p-værdi med anvendelse af BIC-BF, MBF og SD-MBF, og med forskellige á priori sandsynligheder for H_o mellem 5% og 95%. Som det ses, varierer á posteriori sandsynligheden for H_o (ingen forudsigelighed) mellem 0.2% og 83.3%.

Hvordan skal vi fortolke disse resultater? Den klassiske frekventist vil måske hævde, at med så store udsving i P_{H_o} værdierne, er disse værdier ubrugelige, da de ikke giver et klart og 'objektivt' svar på, om aktieafkast er forudsigelige. Men for bayesianeren er det netop pointen: Et sådant klart objektivt svar findes aldrig! Svaret afhænger af den enkelte læsers á priori opfattelse af afkastforudsigelighed (på samme måde som konklusionen på bargæstens korrekte forudsigelse af 10 møntkast afhænger af vores subjektive á priori tro på clairvoyance, uanset p-værdien på 0.001, jf. eksemplet i begyndelsen af afsnit 5). Hvis vi lægger os så tæt op ad den klassiske tankegang som muligt og antager en neutral prior, dvs. $P(H_o) = P(H_1) = 0.5$, og samtidig kræver en meget lav værdi af P_{H_o} for at forkaste H_o (eksempelvis $< 5\%$, svarende til det klassiske 5% signifikansniveau), er det kun MBF der fører til forkastelse af H_o ($P_{H_o} = 0.032$). ($P_{H_o} = 20.8\%$ beregnet med BIC-BF kan evt. også tages som udtryk for relativ stærk evidens imod H_o, givet BIC-BF's konservative natur).

Hvis man derimod har en stærk tro på teorien om efficiente markeder, ifølge hvilken aktieafkast er uforudsigelige, vil ens á priori sandsynlighed for H_o være eksempelvis 80%, hvilket ifølge Tabel 2 giver P_{H_o} værdier på 51.3%, 11.7% og 31.6% for henholdsvis BIC-BF, MBF og SD-MBF. Omvendt hvis man mere tror på *behavioral finance* teorier om inefficiens og irrationalitet, vil ens á priori sandsynlighed for H_o måske være tættere på 20%, hvilket ifølge Tabel 2 giver P_{H_o} værdier på 6.2%, 0.8% og 2.8% for henholdsvis BIC-BF, MBF og SD-MBF. Stikprøvedataene fører for begge personer til en nedjustering af sandsynligheden for uforudsigelighed. Behavioral finance personen bliver bekræftet i sin forhåndsformodning, mens personen, der tror på teorien om efficiente markeder, bliver udfordret på sin tro. Om P_{H_o} værdierne på 11.7% og 31.6% får ham til at skifte tro, er hans personlige valg (de 51.3% for BIC-BF vil under alle omstændigheder ikke føre til en forkastelse af H_o).[18]

[18] Cochrane (2011) konkluderer ikke udfra sine resultater, at aktiemarkedet er irrationelt og inefficient. Han fortolker i stedet den signifikante afkastforudsigelighed som et resultat af rationelt tidsvarierende risikopræmier.

	BIC-BF BIC = -2.669, BF = 0.2633 P_{H_o}	MBF BF = 0.0332 P_{H_o}	SD-MBF BF = 0.1152 P_{H_o}
$P(H_o) = 0.95$	0.833	0.387	0.686
0.80	0.513	0.117	0.316
0.50	0.208	0.032	0.103
0.20	0.062	0.008	0.028
0.05	0.014	0.002	0.006

Note: Á posteriori sandsynligheden P_{H_o} (den bayesianske p-værdi) er beregnet som i formel (7).

Tabel 2: Bayes-faktorer (BF) og tilhørende P_{H_o} for H_o: Ingen afkastforudsigelighed.

5.4 Andre bayesianske metoder til beregning af P_{H_o}.

Jeg har helt bevidst forsøgt at holde beskrivelsen af metoderne illustreret i de foregående afsnit så simpelt som muligt, og med tanke på, at metoderne skal kunne anvendes af empirikere pba. det statistiske output, som de fleste allerede kender. Jeg har beskrevet, hvordan man ret enkelt kan anvende det velkendte T-test for en enkelt parameter (og tilhørede p-værdi) fra den klassiske analyse til at beregne en 'bayesiansk p-værdi', der giver $P(H_o \mid \text{data})$, uden at formulere en eksplicit á priori fordeling for parameteren under H_1.

I praksis vil man naturligvis ofte have brug for i samme test at kunne teste flere parametre i sin model, og man kunne være interesseret i bayesianske p-værdier for andre valg af Bayes-faktorer end BIC-BF, MBF og SD-MBF. Litteraturen indeholder flere forskellige muligheder i så henseende. Leamer (1978, kap. 4) og Zellner and Siow (1979) præsenterer simple formler for beregning af posterior odds ratioer for H_o vs. H_1 i multiple regressionsmodeller, hvor der testes på mere end én af regressionskoefficienterne i samme test, og med anvendelse af 'diffuse' priors (se Connolly (1989, 1991) og Kim and Ji (2015) for anvendelser af disse metoder i empirisk finansiering). Tilsvarende præsenterer Kass and Raftery (1995) formler for BIC-BF når testet indeholder mere end én parameter.[19]

[19] Startz (2014, afsnit 5) indeholder simple formler for P_{H_o} i regressionsmodeller, hvor

6. Metodiske overvejelser om økonomiske modeller.

En grundlæggende præmis for anvendelsen af de klassiske hypotesetests, er, at den statistiske model er velspecificeret og at de fordelingsmæssige antagelser er opfyldt. Hvis data ikke er identisk og uafhængigt normalfordelte, er T-testet i (1) ikke anvendeligt. Modelrammen for metoderne præsenteret i det meste af det foregående har været relativ simpel: Test på én parameter i enten et kontrolleret eksperiment (som eksempelvis at kaste en mønt n gange og teste om sandsynlighedsparameteren er lig med 0.5), eller i en regression á la (9) på ikke-eksperimentelle data. I kontrollerede eksperimenter kan man ofte automatisk sikre, at data opfylder de fordelingsmæssige antagelser. Med ikke-eksperimentelle data, derimod, er dette ikke muligt. I simple modeller som (9) er dette dog typisk ikke noget stort problem. Modellen skal ikke opfattes som en 'strukturel' relation hvor dp_{t-1} 'forårsager' r_t i traditionel økonomisk forstand. (9) er en prediktions-regression, og residualerne behøver ikke være serielt ukorrelerede og homoscedastiske; i tilfælde af autokorrelation og heteroscedasticitet, kan 'robuste' standardafvigelser anvendes ved beregning af t-statistikken.[20]

Problemstillingen bliver væsentligt mere kompliceret når vi bevæger os udenfor den simple modelramme med eksperimentelle data eller simple ikke-strukturelle ligninger som (9). Hvis formålet med den empiriske analyse eksempelvis er at modellere og forstå de kausale sammenhænge mellem finansielle og makroøkonomiske variable, opstår der nye metodemæssige problemer. For de fleste finansielle og makroøkonomiske størrelser er sammenhænge mellem variablene uhyre komplekse med mange baggrundsfaktorer, og komplicerede feed-back mekanismer og dynamiske effekter mellem endogene og eksogene variable, herunder forventningseffekter og laggede effekter som følge af tilpasningsomkostninger m.m. Vi konstruerer økonomiske teoretiske modeller i et forsøg på at forstå hovedmekanismerne i økonomiske individers adfærd, men modellerne kan aldrig blive andet end grove approksimationer til virkeligheden. Ingen model er i stand til at give en komplet og fuldstændig

der testes på én regressionsparameter, θ, og hvor parameteren under H_1 á priori antages at følge enten en uniform fordeling eller en normalfordeling. Formlerne bygger på den klassiske t-statistik og indeholder en parameter, c, der sættes efter hvor bredt man ønsker intervallet omkring θ_o at være under H_1.

[20] Der vil dog typisk være en *finite sample bias* ved OLS estimation af modellen som følge af korrelation mellem modellens residualer og prediktionsvariablens innovationer. Der findes klassiske procedurer til korrektion for denne bias, se Stambaugh (1999).

beskrivelse af virkeligheden, hvilket økonomer naturligvis er fuldt bevidste om.

Ikke desto mindre har det fremherskende paradigme i empirisk analyse af strukturelle økonomiske modeller været at evaluere modellerne på en måde, hvor modellen formuleres som parameterrestriktioner, der kan testes med de klassiske frekvensbaserede hypotesetests. Modellens restriktioner formuleres som en testbar hypotese med beregning af p-værdi, dvs. betinget på, at H_o (modellen) er sand (som beskrevet i de foregående afsnit). Et velkendt eksempel er tests af overidentificerende restriktioner i rationelle forventningsmodeller, eksempelvis Hansen and Singleton (1983).[21]

Allerede Kydland and Prescott (1982) var skeptiske overfor denne tankegang. Deres Nobelprisvindende model var aldrig tænkt som en præcis beskrivelse af virkeligheden. De skriver: "*We choose not to test our model . . . this most likely would have resulted in the model being rejected, given the measurement problems and the abstract nature of the model*" (p.1360). I stedet har relativt uformelle kalibrerings- og simulationsteknikker traditionelt været anvendt til at evaluere *Real Business Cycle* (RBC) og *Dynamic Stochastic Generel Equilibrium* (DSGE) modeller. Men de senere år har den bayesianske tilgang vundet frem indenfor DSGE modellering, i et mere formaliseret forsøg på at adressere den kritik, der berettiget blev rejst mod de uformelle evalueringsteknikker i den tidlige RBC litteratur.

Som fremhævet af Schorfheide (2000) og Fernandez-Villaverde and Rubio-Ramirez (2004), er bayesianske metoder specielt velegnede til evaluering af fejlspecificerede modeller. DSGE modeller er per konstruktion fejlspecificerede ("*Since a dynamic equilibrium economy is an artificial construction, these models will* always *be false.*" (Fernández-Villaverde and Rubio-Ramírez, 2004, p.153)). Dette i sig selv gør det problematisk at evaluere en sådan model vha. klassiske frekvensbaserede hypotesetests. "*The potential misspecification of the candidate models poses a conceptual difficulty for the design of econometric procedures. The inference problem is not to determine whether a particular model ... is 'true'. Instead, the goal is to determine which model summarizes the regular features of the data more accurately.*" "*Many evaluation techniques that were previously proposed in the literature are based on p-values for various sample characteristics of the data. ... How-*

[21] Engsted (2002) giver et survey over metoder til at evaluere simple rationelle forventningsmodeller, herunder metoder, der ikke bygger på klassiske hypotesetests og som tillader fejlspecifikation i modellerne. Ingen af de gennemgåede metoder har dog et bayesiansk tilsnit.

ever, p-values are not designed for model comparisons. Moreover, in cases where it is believed that the structural models are severely misspecified it is implausible to use sampling distributions as a benchmark, that were derived from misspecified models." (Schorfheide, 2000, pp.645-646).

Den bayesianske tilgang, derimod, tillader fejlspecifikation og muliggør sammenligning af to (eller flere) modeller, der begge grundlæggende er fejlspecificerede. Den bayesianske *posterior odds ratio* i (4) forudsætter ikke, at en af modellerne er 'sand', men fortæller blot, hvilken af de to modeller der er mest '*likely*'. "*Bayesian inference builds on the insight that models are false and is ready to deal with this issue in a natural way. Estimation moves from being a process of discovery of some "true" value of a parameter to being a selection device in the parameter space that maximizes our ability to use the model as a language in which to express the regular features of the data.*" (Fernández-Villaverde and Rubio-Ramírez, 2004, p.154). Den bayesianske tilgang er blevet den fremherskende ved estimation og evaluering af DSGE modeller, eksemplificeret af den kanoniske 'Smets-Wouters model' (Smets and Wouters, 2003 og 2007), der er blevet en af de mest citerede og anvendte makromodeller i nyere tid.

Der har været forslag fremme om helt at droppe anvendelsen af klassiske hypotesetests indenfor samfundsvidenskab (og andre videnskaber), se eksempelvis Ziliak and McCloskey (2008). Det finder jeg vil være et unødigt drastisk skridt at tage. Jeg har tidligere (Engsted, 2009) argumenteret for, at de klassiske tests fortsat er nyttige ved specifikation af rene statistiske modeller og kan spille en rolle i analyser af strukturelle økonomiske modeller. I mange sammenhænge indgår statistiske modeller (eksempelvis VAR modeller) som input i økonomiske modeller. De klassiste tests kan her være et nyttigt redskab til at opnå en velspecificeret og parsimon statistisk model. Men når denne model efterfølgende anvendes i kvantificeringen af den økonomiske model, bør evalueringen af denne ikke ske med udgangspunkt i klassiske hypotesetests. Her vil bayesianske metoder efter min opfattelse være mere velegnede.

7. Konklusion.

Ingen videnskabelig disciplin er objektiv eller værdifri, og statistisk analyse indeholder mange subjektive elementer. Det gælder uanset om vi taler om de klassiske frekvensbaserede eller de bayesianske analyser. I de klassiske hypotesetest viser subjektiviteten sig især i valget af et (arbitrært)

signifikansniveau. I den bayesianske analyse er det valget af prior der er subjektivt.

Den klassiske analyses 'p-værdi' giver sandsynligheden for teststatistikken, givet H_o, dvs. $P(\text{data} \mid H_o)$. Den bayesianske analyse giver derimod $P(H_o \mid \text{data})$, der for en samfundsforsker mere naturligt relaterer til de modeller vi arbejder med. Ikke mindst hvis man anerkender, at vores modeller er nyttige, men samtidig abstrakte og grove approksimationer til virkeligheden.

Á posteriori sandsynligheden for H_o - den bayesianske p-værdi, $P(H_o \mid \text{data})$ - er generelt væsentlig højere end den klassiske p-værdi, hvilket indikerer, at den klassiske frekvensbaserede analysemetode givetvis fører til mange 'falsk positive' resultater.

Samfundsforskere bør - ligeså naturligt som vi anvender de klassiske frekvensbaserede metoder - anvende bayesianske metoder i vores analyser. Og bayesianske metoder bør spille en lige så vigtig rolle i statistik- og økonometriundervisningen på universiteterne, som de klassiske metoder spiller og altid har spillet.

8. Appendix.

Analytisk beregning af P_{H_o} i møntkasteksemplet i afsnit 4.3:

For H_o: $\theta = 0.5$ og H_1: $\theta \neq 0.5$, bliver formel (6) til

$$P_{H_o} = \frac{P(\text{data} \mid H_o)}{P(\text{data} \mid H_o) + \int P(\text{data} \mid H_1) f(\theta_1) d\theta_1}. \tag{A1}$$

Med en uniform prior for θ_1 i intervallet (0, 1), dvs. $f(\theta_1) = 1$, bliver integralet i (A1) til

$$\begin{aligned} \int P(\text{data} \mid H_1) f(\theta_1) d\theta_1 &= \int_0^1 \left(\binom{n}{m} \theta_1^m (1-\theta_1)^{n-m} \right) d\theta_1 \\ &= \binom{n}{m} \int_0^1 (\theta_1^m (1-\theta_1)^{n-m}) d\theta_1 \\ &= \frac{n!}{m!(n-m)!} \cdot \frac{m!(n-m)!}{(n+1)!} = \frac{1}{n+1}, \end{aligned} \tag{A2}$$

hvor det bemærkes, at $P(\text{data} \mid H_1)$ er en binomialsandsynlighed, og $\int_0^1 (\theta_1^m (1-\theta_1)^{n-m}) d\theta_1$ er den såkaldte *beta-funktion* med værdien $\frac{m!(n-m)!}{(n+1)!}$ (jf. Berry and Lindgren, 1996, p.355).

Hermed fås

$$P_{H_o} = \frac{P(\text{data} \mid H_o)}{P(\text{data} \mid H_o) + \frac{1}{n+1}}. \tag{A3}$$

I eksemplet med $n = 32$ og $m = 21$, fås $P(\text{data} \mid H_o) = 0.03$ (binomialsandsynlighed for 21 'succes' i 32 forsøg med sandsynlighedsparameter = 0.5), hvorved P_{H_o} ifl. (A3) beregnes til

$$P_{H_o} = \frac{0.03}{0.03 + \frac{1}{33}} = 0.50.$$

9. Referencer.

Angrist, J., P. Azoulay, G. Ellison, R. Hill, and S.F. Lu (2017): Economic research evolves: Fields and styles. *American Economic Review: Papers and Proceedings* 107, 293-297.

Bayarri, M.J. and J.O. Berger (1998): Quantifying surprise in the data and model verification. In: Bernardo, J.M., J.O. Berger, A.P David, and A.F.M. Smith (eds.), *Bayesian Statistics*, Vol. 6. Oxford University Press, Oxford.

Berger, J.O. and M. Delampady (1987): Testing precise hypotheses. *Statistical Science* 2, 317-352.

Berger, J.O. and T. Sellke (1987): Testing a point null hypothesis: The irreconcilability of P values and evidence. *Journal of the American Statistical Association* 82, 112-122.

Berry, D.A. and B.W. Lindgren (1996): *Statistics: Theory and Methods* (2nd edition). Duxbury Press.

Børsen (2015): Elbæks parti er meget tæt på spærregrænsen. 2. marts 2015, side 18.

Campbell, J.Y. and R.J. Shiller (1988): The dividend-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies* 1, 195-228.

Cochrane, J.H. (2011): Presidential Address: Discount rates. *Journal of Finance* 66, 1047-1108.

Connolly, R.A. (1989): An examination of the robustness of the weekend effect. *Journal of Financial and Quantitative Analysis* 24, 133-169.

Connolly, R.A. (1991): A posterior odds analysis of the weekend effect. *Journal of Econometrics* 49, 51-104.

Edwards, W., H. Lindman and L.J. Savage (1963): Bayesian statistical inference for psychological research. *Psychological Review* 70, 193-242.

Engsted, T. (2002): Measures of fit for rational expectations models. *Journal of Economic Surveys* 16, 301-355.

Engsted, T. (2009): Statistical vs. economic significance in economics and econometrics: Further comments on McCloskey and Ziliak. *Journal of Economic Methodology* 16, 393-408.

Fernandez-Villaverde, J., and J.F. Rubio-Ramirez (2004): Comparing dynamic equilibrium models to data: A Bayesian approach. *Journal of Econometrics* 123, 153-187.

Fisher, R.A. (1925): *Statistical Methods for Reseach Workers.* Oliver and Boyd Ltd., Edinburgh.

Greenhouse, J.B. (2012): On becoming a Bayesian. Early correspondances between J. Cornfield and L.J. Savage. *Statistics in Medicine* 31, 2782-2790.

Hansen, L.P. and K.J. Singleton (1983): Stochastic consumption, risk aversion, and the temporal behavior of asset returns. *Journal of Political Economy* 91, 249-265.

Harvey, C.R. (2017): Presidential address: The scientific outlook in financial economics. *Journal of Finance* 72, 1399-1440.

Harvey, C.R., Y. Liu and H. Zhu (2016): ... and the cross-section of expected returns. *Review of Financial Studies* 29, 5-68.

Ioannidis, J.P.A. (2005): Why most published research findings are false. *PLoS Medicine* 2, e124.

Kass, R.E. and A.E. Raftery (1995): Bayes factors. *Journal of the American Statistical Association* 90, 773-795.

Keuzenkamp, H.A. and J.R. Magnus (1995): On tests and significance in econometrics. *Journal of Econometrics* 67, 5-24.

Kim, J.H. and I. Choi (2017): Unit roots in economic and financial time series: A re-evaluation at the decision-based significance levels. *Econometrics* 5, 41.

Kim, J.H. and P.I. Ji (2015): Significance testing in empirical finance: A critical review and assessment. *Journal of Empirical Finance* 34, 1-14.

Kydland, F.E. and E.C. Prescott (1982): Time to build and aggregate fluctuations. *Econometrica* 50, 1345-1370.

Leamer, E.E. (1978): *Specification Searchers: Ad Hoc Inference with Non Experimential Data*. John Wiley & Sons.

Neyman, J. and E.S. Pearson (1933): On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society* Series A 231, 289-337.

Politiken (2014): Overblik. 30. april, 2014, side 9.

Raftery, A.E. (1999): Bayes factors and BIC. *Sociological Methods & Research* 27, 411-427.

Schorfheide, F. (2000): Loss function-based evaluation of DSGE models. *Journal of Applied Econometrics* 15, 645-670.

Sellke, T., M.J. Bayarri, and J.O. Berger (2001): Calibration of p values for testing precise null hypotheses. *The American Statistician* 55, 62-71.

Smets, F. and R. Wouters (2003): An estimated dynamic stochastic general equilibrium model of the Euro area. *Journal of the European Economic Association* 1, 1123-1175.

Smets, F. and R. Wouters (2007): Shocks and frictions in US business cycles: A Bayesian DSGE approach. *American Economic Review* 97, 586-606.

Stambaugh, R.F. (1999): Predictive regressions. *Journal of Financial Economics* 54, 375-421.

Startz, R. (2014): Choosing the more likely hypothesis. *Foundations and Trends in Econometrics* 7, 119-189.

Wasserstein, L.R. and N.A. Lazar (2016): The ASA's statement on p-values: Context, process, and purpose. *American Statistician* 70, 129-133.

Zellner, A. (1971): *An Introduction to Bayesian Inference in Econometrics*. John Wiley & Sons.

Zellner, A. and A. Siow (1979): Posterior odds ratios for selected regression hypotheses. In: Bernardo, DeGroot, Lindley, and Smith (eds.): *Bayesian Statistics: Proceedings of the first International Meeting Held in Valencia, Spain*. University Press Valencia, pp. 585-603.

Ziliak, S.T. and D.N. McCloskey (2008): *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Ann Arbor: The University of Michigan Press.

Economics Working Papers

2017-08: Marianne Simonsen, Lars Skipper and Niels Skipper: Piling Pills? Forward-Looking Behavior and Stockpiling of Prescription Drugs

2017-09: Federico Ciliberto and Ina C. Jäkel: Superstar Exporters: An Empirical Investigation of Strategic Interactions in Danish Export Markets

2017-10: Anna Piil Damm, Britt Østergaard Larsen, Helena Skyt Nielsen and Marianne Simonsen: Lowering the minimum age of criminal responsibility: Consequences for juvenile crime and education

2017-11: Erik Strøjer Madsen: Branding and Performance in the Global Beer Market

2017-12: Yao Amber Li, Valerie Smeets and Frederic Warzynski: Processing Trade, Productivity and Prices: Evidence from a Chinese Production Survey

2017-13: Jesper Bagger, Espen R. Moen and Rune M. Vejlin: Optimal Taxation with On-the-Job Search

2018-01: Eva Rye Johansen, Helena Skyt Nielsen and Mette Verner: Long-term Consequences of Early Parenthood

2018-02: Ritwik Banerjee, Nabanita Datta Gupta and Marie Claire Villeval: Self Confidence Spillovers and Motivated Beliefs

2018-03: Emmanuele Bobbio and Henning Bunzel: The Danish Matched Employer-Employee Data

2018-04: Martin Paldam: The strategies of economic research - An empirical study

2018-05: Ingo Geishecker, Philipp J.H. Schröder, and Allan Sørensen: One-off Export Events

2018-06: Jesper Bagger, Mads Hejlesen, Kazuhiko Sumiya and Rune Vejlin: Income Taxation and the Equilibrium Allocation of Labor

2018-07: Tom Engsted: Frekvensbaserede versus bayesianske metoder i empirisk økonomi