Hypotesetest – helt så let er det desværre ikke

Jan B. Sørensen, Aalborg Studenterkursus

Indførelsen af χ^2–hypotesetest på stx så umiddelbart meget simpel ud. Selve beregningerne er lette, mens teorien bag er så svær, at vi i gymnasiet alligevel ikke realistisk kan behandle den. χ^2–test har også den fordel, at der ikke stilles krav om, at data fx er normalfordelte (man siger, at testet er ikke–parametrisk), så modelkontrol er unødvendig. Endeligt kan kvalitative (kategoriske) såvel som kvantitative (numeriske) data benyttes, idet sidstnævnte blot skal grupperes, så også i den henseende har testene bred anvendelse. χ^2–test er kort og godt meget lette at finde anvendelser for, og CAS værktøjerne leverer på ingen tid alle beregningerne.

Men, der er alligevel nogle gråzoner, som man bør være opmærksom på som lærer og måske også som elev. I dette indlæg vil jeg kort præsentere nogle af disse og opfordre til, at man ude på skolerne tager en debat i faggruppen eller måske endda rekvirerer Matematiklærerforeningens kursus om hypotesetest, samt at man holder et vågent øje med samfundsfags–, biologi–, og andre lærere, når de benytter χ^2–test.

Statistisk signifikans kan fremprovokeres. I april var der en meget anbefalelsesværdig artikel i Ingeniøren. Artiklen kan læses her: ing.dk/artikel/128347. Meget kort er artiklens budskab, at man i god eller i ond tro nærmest altid kan producere et signifikant resultat, hvis man bare tester uhæmmet.

Hvis man fx laver χ^2–test på en spørgeskemaundersøgelse med 10 forskellige spørgsmål, hver på 5% signifikansniveau, så bliver sandsynligheden for at forkaste mindst ét af disse cirka 40% (under antagelse af, at testene er uafhængige af hinanden), selvom alle 10 reelt var sande. Udvid så tanken til 50 spørgsmål med tilhørende test, og forkastede hypoteser bliver pludseligt ikke særligt interessante, som de ellers burde have været. Man kan også forestille sig en elev, der brændende ønsker at sige noget interessant om en meningsmåling, uanset hvad. Eleven kigger på meningsmålingen og laver så et test specifikt på det parti, der ser ud til at have haft den mest markante ændring i stikprøven. I en vis udstrækning er det samme situation som de 10 spørgsmål i spørgeskemaet. Eleven laver godt nok ikke eksplicit et test på hver af partierne, man ved selektivt at gå efter det parti med størst ændring i stikprøven, er det jo samme problematik.

Hvis man ser på de vejledende opgaver i χ^2, sker faktisk netop dette i opgaven, hvor SF udvælges til test i sidste delspørgsmål. Jeg snakkede på et tidspunkt med Susanne Christensen, der er statistiker på Aalborg Universitet og forfatter til den note, som blev lavet som oplæg til indførelsen af χ^2. Ifølge hende er det en gråzone, hvis man ansporet af data i en stikprøve foretager et test på samme stikprøve. Nogle statistikere gør dette uden at blinke, mens andre hellere vil indsamle en ny stikprøve og lave testet der, men hvor det er upraktisk/umuligt, laver de testet på samme stikprøve, uanset problematikken. Jeg synes dog, at man som minimum bør overveje den mulige fejl, hver gang man gør dette, og hvor det er muligt i stedet teste i en frisk stikprøve.

For stor stikprøve i endelig population

Ingen er vel i tvivl om, at en for lille stikprøve giver problemer, men for stor en stikprøve kan også give problemer, især i en endelig population. Hvis man tager en stikprøve på 80 personer ud af alle danskere, er alt godt, da der stadig er plads til at have taget stikprøven på mange andre måder. Hvis man derimod tager en stikprøve på 80 personer på en arbejdsplads med 100 personer, er man i problemer. Tester man fx en hypotese om, at 75% af de ansatte er glade for arbejdspladsen og observerer 53 glade blandt de 80 i stikprøven, så er det jo ikke muligt at nå op på 75%, selv hvis alle de resterende 20 er glade, men et χ^2–GoF–test forkaster ikke hypotesen om de 75% (p–værdi på 7%). Dette er naturligvis et overdrevet eksempel, men det viser, at stikprøven ikke må være for stor en andel af population. Jeg har ikke set en officiel tommelfingerregel, men uofficielt har jeg af en professor i statistik (tidligere KU), fået at vide, at stikprøven ikke må være mere end 1/20 af populationen, hvilket passer godt med min egen mavefornemmelse, så i mangel af bedre viden vil jeg benytte dette forhold. Hvis denne tommelfingerregel overskrides, bliver den tilbageværende varians så markant mindre, at det fastsatte signifikansniveau ganske enkelt ikke længere passer. Man kan så evt. korrigere for denne manglende varians, men det er ud over denne artikel. Endnu værre bliver det naturligvis, hvis man spørger alle 100 af 100 medarbejdere, da der så ikke er noget ukendt tilbage, og opgaven derfor reduceres til deskriptiv statistik.

Hvis man ser på en opgave med χ^2, som landets skriftlige censorer skulle forholde sig til på et møde i foråret, indeholder den faktisk netop dette problem. Her får man en procentvis fordeling mellem hjemmesejr, uafgjort og udesejr i en sæson i en engelsk fodboldliga. Man får desuden en stikprøve fra en igangværende sæson, hvor man har observeret frem til medio november. Opgaven uden data lyder således:

> *Opstil en nulhypotese, og undersøg på et 5% signifikansniveau, om resultatfordelingen pr. 14/11 for sæsonen 2010/2011 følger samme fordeling som resultatfordelingen for sæsonen 2009/2010.*

For det første er det nok et problem, at langt fra alle gymnasieelever ved, at sæsonen løber fra sommer til sommer og har 380 kampe i alt (i Premier League, som jeg antager data er fra), så elever uden denne viden vil få svært ved reelt at vurdere, om stikprøven er for stor. For det andet er en stikprøve på cirka 1/3 af sæsonen klart for stor en andel efter min bedste overbevisning. Endeligt er formuleringen faktisk sådan, at stikprøven = populationen, hvis man læser op-

gaven ordret. Der står *om resultatfordelingen pr. 14/11 for sæsonen 2010/2011*. Det er givet ikke mening, at populationen kun skal være kampene frem til medio november, så formuleringen burde nok ændres, så det fremgår, at stikprøven er disse kampe, mens populationen er hele sæsonen 2010–11, fx ved blot at slette ordene pr. 14/11.

Potentiel (uendelig) population

I mange tilfælde kommer man uden om førnævnte problem ved at se data som en stikprøve fra en potentiel større, evt. uendelig population, i modsætning til en konkret realiseret population. Et klokkerent eksempel på dette er kast med en terning, hvor man ser de faktiske kast som en stikprøve af alle potentielle kast med terningen. Dette er også ofte metoden i forbindelse med sygdomme og lignende. Hvis Rigshospitalet derfor undersøger alle danskere med en bestemt sjælden sygdom, så anses disse personer som en stikprøve ud af alle personer, der potentielt kan have eller få sygdommen. Dette giver ofte rigtig god mening, hvis man ønsker at udtale sig om den potentielle population, frem for den konkrete, synlige population. Efter min mening skal man dog være varsom med blot blindt at benytte denne metode. Fx i fodboldopgaven før giver det dårlig mening, da undersøgelsen handler om den konkrete sæson, så at tale om disse kampe som en stikprøve fra en større (fiktiv) mængde af kampe giver næppe mening. Det samme vil ofte være tilfældet, hvis en virksomhed undersøger sine medarbejdere. Der vil virksomheden oftest også være interesserede i de konkrete medarbejdere, frem for en fiktiv større mængde.

Eksempel: Givet fordelingen af blodtyper i hele den danske befolkning laver en klasse en opgørelse over fordelingen af blodtyper for eleverne i klassen. Stikprøven er klar, men hvad er populationen? Hvis eleverne kun ønsker at udtale sig om klassen selv, så er stikprøve = population, og test er nonsens. Hvis klassen derimod ønsker at udtale sig om alle skolens elever og ser sig selv som en repræsentativ stikprøve (nok rimeligt, når der er tale om blodtyper), så fungerer det fint, hvis skolen er stor. Hvis klassen ikke ønsker at udtale sig om en konkret population, men derimod vil se på en potentiel/fiktiv population, som de er en delmængde af, så bliver det efter min mening lidt søgt. Hvad skulle potentialet være? Alle de elever, som kunne have gået i klassen? For mig giver det i hvert fald ingen mening, så jeg vil opfordre til, at man kun benytter opgaver med potentielle populationer, hvor man har et veldefineret formål, så populationen bliver meningsfuld. Faktisk skrev Susanne Christensen i forbindelse med en opgaveformulering, hvor man skulle teste om "fordelingen af blodtyper i klassen er den samme som i hele den danske befolkning", at *'Præcist det punkt har jeg diskuteret med mange gymnasielærer også. Det er en fælde, man ofte går i ved testene for fordeling'*, så i hvert fald nogle statistikere mener, at man skal træde varsomt.

For stor stikprøve i en uendelig population

Umiddelbart skulle man tro, at en større stikprøve så uden undtagelse vil være bedre, hvis populationen er uendelig. Dette er dog en sandhed med forbehold. Antag, at man tester om 60% af en population er positive, men at det reelt er 60,04%, der er positive. I mange praktiske sammenhænge vil man nok sige, at så er 60% sandt, fordi det rigtige tal er meget tæt på og afrundes til 60,0%, men reelt er det jo falsk. Hvis man benytter en moderat stikprøve, vil den stort set ikke kunne kende forskel på 60% og 60,04%, men hvis man benytter en meget stor stikprøve, vil den kunne "opdage" den lille forskel. Man bør derfor overveje at sætte signifikansniveauet lavere, hvis man har ekstremt store stikprøve, hvis man ikke vil forkaste pga. meget små forskelle. I praksis er dette dog sjældent et problem, da man som oftest ikke har så store stikprøver, men overvej lige problemet, hvis det skulle ske.

Kort kommentering af de vejledende χ^2–opgaver

Der er udarbejde en række vejledende opgaver med χ^2–test. Opgaverne ligger bl.a. elektronisk (Google og du skal finde), så når jeg henviser til opgavenumre, er det til den elektroniske version. Jeg har ladet mig fortælle, at de i papirudgaven har andre numre, men har ikke selv opgaverne på papir.

1 Hjertecentre

En meget konkret læsning af opgaven kunne give det indtryk, at stikprøve = population, da data for alle gennemførte operationer er med. Her skal man tænke population mere abstrakt, nemlig som de potentielle operationer, der kunne være foretaget og evt. kan blive foretaget i fremtiden. Om stikprøven så er repræsentativ, især mht. fremtidige operationer, kan (og bør) være til debat. Opgaven lægger da også helt rigtigt op til en diskussion af skjulte variable. Tilsvarende kommentarer gælder til opgave 5 om hjerte– og lungeoperationer.

2 X–køb ing

Fin opgave. Populationen er godt nok ikke helt veldefineret, men det er let at finde på en plausibel population. Med formuleringen X–køb ing indikeres også en population af en vis størrelse, så der skulle ikke være problemer med, at stikprøven udgør for stor en andel. Tilsvarende kommentarer gælder til opgave 4 om medicin og opgave 6 om rygevaner.

3 Mus

Der er klart under 5 i en af cellerne med forventede værdier, så opgaven må ikke løses vha. χ^2–test! Opgaven kunne løses vha. Fishers eksakte test (der bygger på hypergeometriske fordelinger) eller simulering, men det er næppe tanken. Tallene i opgaven bør derfor laves om, hvorefter det er en rigtig god opgave.

7 Drikkervaner

"... undersøge, om drikkevaner er uafhængig af køn" - men spørger kun elever. Dette bør give anledning til skarp kritik.

Det er ganske enkelt for uklart, hvilken population der testes i – alle, alle elever, alle elever på en bestemt type uddannelse, alle elever på en bestemt skole eller noget helt femte. Afhængigt af populationen, kan stikprøven også risikere at være for stor en andel.

8 Legetøjsbolde

Hvis man er meget krakilsk, er spørgsmålet formuleret uheldigt. Her hentyder jeg til verbet "kan". Eftersom ingen af frekvenserne i fordelingen er 0%, *kan* enhver stikprøve stamme fra producenten. I stedet burde der måske stå "Undersøg, om det på 5% signifikansniveau kan forkastes, at sendingen stammer fra storproducenten". Derudover er opgaven god.

9 Blomster og 10 Klager

Intet at bemærke.

11 Folketingsvalg

Her skal man være opmærksom på, at man aldrig må foretage test på frekvenser, men skal omregne til hyppigheder, da en skalering af data betyder meget (2 6'ere i 6 kast med en terning er ok, mens 200 6'ere i 600 kast er meget suspekt). Man skal derfor omregne både fordelingen og stikprøven fra frekvenser til hyppigheder. I den forbindelse skal fordelingen omregnes til hyppigheder med decimaler, mens stikprøven optimalt set bør omregnes til hele tal, hvis dette er muligt. I tilfælde, hvor det ikke er muligt, må man nøjes med også at omregne stikprøven til hyppigheder med decimaler. Frekvenserne summerer til 100,1 %, men med et total antal på 968 er de angivne frekvenser ganske enkelt ikke mulige, så der må være en trykfejl i opgaven, enten mht. det totale antal eller mht. frekvenserne. I del c af opgaven ryger vi desuden ind i den gråzone, jeg omtalte tidligere, når SF på baggrund af stikprøvens data udvælges til yderligere hypotesetest.

Kort kommentering af de stillede eksamensopgaver den 25/5 og 31/5

Til eksamen på B–niveau den 25. maj 2012 blev der i opgave 12 stillet en Goodness–of–fit opgave med χ^2. Sværhedsgraden af den stillede opgave er, helt som forventet, meget lav, da man meget fornuftigt har valgt at starte med en let opgave, hvor CAS–værktøj typisk løser hele opgaven uden behov for nærmere overvejelser. Læst meget bogstaveligt har opgaven et problem med, at stikprøven på de 950 patienter også er populationen, men som for flere af de vejledende opgaver skal man tænke på populationen som potentielle patienter – enten som et potentiale over de kommende år, eller som potentiale i form af, at andre kunne have benyttet klinikken. Hvis klinikken er specialiseret i en bestemt type patienter, kan det så blive en test for, om denne type patienter har samme fordeling i blodtyper som hele befolkningen.

Til eksamen på B–niveau den 31. maj 2012 blev der i opgave 11 ligeledes stillet en opgave i Goodness–of–fit, denne gang med en Megafon–undersøgelse om vælgertilslutning til de politiske partier. Igen en meget let opgave, og denne gang helt uden problemer mht. population, størrelse af stikprøve eller andet.

Jeg mener som sagt, at det er meget fornuftigt at starte med så lette opgaver, men synes dog på sigt at sværhedsgraden og især abstraktionen godt kan hæves. Især kan jeg forestille mig, at der med fordelt stilles opgaver, hvor der spørges til, hvad der er stikprøve og hvad der er population (som i opgave 11 i de vejledende opgaver). Efter min mening er det svært at give vores elever en reel viden om, hvad hypotesetest kan bruges til, hvis der ikke bliver fokus på netop dette, og skriftlig eksamen er klart det stærkeste kort, når der skal skabes fokus. På samme måde kan der også spørges ind til skjulte variable, og måske endda konstrueres opgaver, hvor Simpsons paradoks indgår. Man kan også lave opgaver med flere kategorier, hvor der er nødvendigt at slå kategorier samme for at lave testet, så denne regel ikke overses.

Kommentarer og spørgsmål modtages meget gerne pr. mail eller endnu bedre i form af indlæg i dette blad. Efter min mening er hele emnet hypotesetest én stor gråzone, så jeg har givet overset noget, som andre måske kan belyse.