METODE & FORSKNINGSDESIGN
https://journals.aau.dk/index.php/mf

Nr. 4 · 2022

Introduktion til statistisk inferens eksemplificeret med sikkerheds-intervaller og hypotesetest

FORMIDLINGSARTIKEL

Henrik Lauridsen Lolle
Lektor, Institut for Politik og Samfund, Aalborg Universitet (lolle@dps.aau.dk)

Resumé:

Artiklen er tænkt som en ikke-statistiktung indføring i principperne for statistisk inferens, dvs. hvordan man på baggrund af en stikprøve kan sige noget om forhold i den bagvedliggende population med en vis statistisk sikkerhed. Der fokuseres først og fremmest på de rent logiske principper, men samtidig hermed bliver den bagvedliggende teoretiske statistik forklaret så tilpas fyldestgørende, at man efter endt læsning forhåbentlig ikke sidder tilbage med en fornemmelse af at mangle noget, med mindre selvfølgelig at man er specielt interesseret i det strengt statistiske. For de fleste samfundsvidenskabelige uddannelser vurderer jeg således, at den vil have et passende statistisk niveau og under alle omstændigheder som en introduktion, der kan bygges videre på. Men artiklen er tænkt sådan, at man som læser efterfølgende skulle kunne kaste sig over litteratur om statistiske metoder som fx stikprøveteori, korrelationsanalyse og lineær regression. Introduktionen i statistisk inferens bliver eksemplificeret med sikkerhedsintervalestimering og statistisk hypotesetest for henholdsvis gennemsnit og proportioner.

ISSN: 2245-3083

1. Introduktion

Der findes overordnet to typer af statistik, *deskriptiv* og *analytisk*. Den deskriptive statistik, som også kaldes for *beskrivende* statistik, beskriver ved hjælp af grafer, tabeller og statistiske størrelser, som fx gennemsnit og standardafvigelse, forhold i et datasæt, hvor datasættet enten kan være i form af en stikprøve eller populationstal. Det kan gøres med mere eller mindre avancerede statistiske metoder, men hovedformålet er, statistisk set, ren beskrivelse. Anderledes forholder det sig med den analytiske statistik, også kaldet *inferential* statistik eller *teoretisk* statistik. At inferere betyder at foretage slutninger på en ukendt situation på baggrund af erfaring og viden fra en kendt situation. I sammenhæng med analytisk statistik betyder det mere præcist, at man pga. dels en konkret viden om data fra en *stikprøve*, dels en teoretisk viden om, hvordan stikprøver typisk arter sig, konkluderer om forhold i den *population*, hvorfra stikprøven er udtrukket. Der er selvfølgelig usikkerhed forbundet med sådan en slutning, og derfor konkluderer man også med en vis *statistisk usikkerhed*. I artiklen gennemgås nogle helt basale grundlag for den teoretiske statistik samt forskellige simple situationer, hvor man bruger den teoretiske statistik i praksis. Fordi der netop gennemgås basalt statistisk stof, har teksten stort set ikke nogen referencer. Der ses dog hos forskellige forfattere lidt forskellige notationer i formler, og her følger jeg næsten fuldt ud Alan Agresti (2018)[1], hvortil jeg også referer enkelte gange. I afsnit 5 her jeg lidt afvigelse i notationen ift. Agresti. De betegnelser, jeg benytter for størrelser som gennemsnit, standardafvigelse osv., fremgår af Bilag 1. Heraf fremgår endvidere formler for netop disse to størrelser, sådan som de beregnes ud fra en stikprøve.

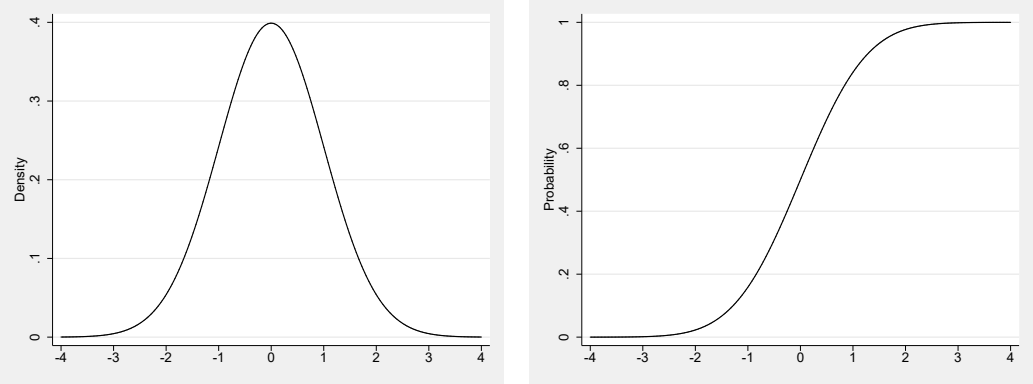
Noget af det allermest grundlæggende for den teoretiske statistik er *normalfordelinger*. Jeg vil herunder først ganske kort introducere normalfordelinger og dernæst forklare, hvorfor de har så vigtig en funktion i analytisk statistik. Normalfordelinger er en familie af klokkeformede kurver, der fordeler sig symmetrisk omkring deres gennemsnit. I Figur 1A er vist en såkaldt *standardnormalfordeling* med et gennemsnit på 0 og en standardafvigelse på 1. Af figuren fremgår det, hvordan flest observationer ligger tæt på gennemsnittet, og hvor der bliver færre og færre observationer, jo længere væk fra gennemsnittet man går i den ene eller anden retning. Læg imidlertid mærke til, at y-aksen i figuren ikke angiver antal observationer. Den viste fordeling er en teoretisk *sandsynlighedsfordeling*, hvor arealet under kurven angiver sandsynlighed, og kurven kaldes nogle gange for en *tæthedskurve*, på engelsk en *density curve*. Halerne i hver side af normalfordelingen går mod 0 på y-aksen, når x går mod $\pm\infty$, men de når aldrig helt ned på 0. Det samlede areal under kurven giver 1, og det betyder blot, at et tilfældigt valgt tal fra fordelingen har en sandsynlighed på 1, dvs. 100 pct., for at ligge mellem $-\infty$ og $+\infty$, hvilket jo ikke er overraskende. Det er imidlertid også tydeligt ud fra figuren, at sandsynligheden er meget tæt på 1 for, at dét tal, man trækker, ligger mellem

[1] *Statistical Methods for the Social Sciences*. Pearson. 5th (global) edition.

-3 og +3. Samme standardnormalfordeling er i øvrigt også vist i Figur 1B, blot i en udgave der viser den *kumulerede* sandsynlighed på y-aksen. Der findes en matematisk formel til bestemmelse af normalfordelinger ud fra oplysning om alene gennemsnit og standardafvigelse, som imidlertid ikke er væsentlig i den her forbindelse. Men tilbage til spørgsmålet om, hvorfor er den slags fordelinger egentlig så vigtige for den teoretiske statistik. Det hænger sammen med et fund, statistikere i sin tid gjorde angående den måde, som stikprøver "opfører" sig på.

Figur 1 Standard normalfordeling

A. Tæthedskurve *B. Kumulativ sandsynlighedskurve*



Lad os gøre det simpelt og sige, at man blot er interesseret i at vide noget om gennemsnitsværdien i en population på en intervalskaleret variabel. Statistikerne fandt ud af, at hvis man *simpelt tilfældigt* trækker en meget lang række af stikprøver, fx med 500 observationer i hver, fra én og samme population, hvor man i alle stikprøverne måler på den samme intervalskalerede variabel og beregner gennemsnit på den, så vil disse mange stikprøvers *gennemsnitsværdier* på variablen fordele sig med noget, der til forveksling ligner en normalfordeling. Læg mærke til, at det altså ikke er fordelingen i populationen eller fordelingen i den enkelte stikprøve, der er tale om her, men derimod en fordeling af de enkelte stikprøvers gennemsnit på variablen. Gennemsnitsværdierne kan betragtes som en variabel på et aggregeret niveau, og det er disse mange gennemsnit, der fordeler sig meget lig en normalfordeling. Det gælder vel og mærke, uanset hvordan pågældende variabel er fordelt i populationen og i den enkelte stikprøve. Sådan en type af fordelinger af stikprøve*statistikker*, altså fx fordelinger af gennemsnit, kaldes for en *samplingfordeling*. Men det gælder kun ved store stikprøver, at samplingfordelingen af gennemsnit ligner en normalfordeling, uanset hvordan variablen er fordelt i populationen. *Hvis* variablen, som man måler, *er* normalfordelt i populationen, vil samplingfordelingen af gennemsnit altid være normalfordelt uanset stikprøvestørrelse, men ved andre typer af fordeling i populationen forholder det sig kun sådan, at des større stikprøver der udtrækkes, des mere *tilnærmer* sampling-fordelingen sig en normalfordeling. Og den lov, der siger, at stikprøvegennemsnittene

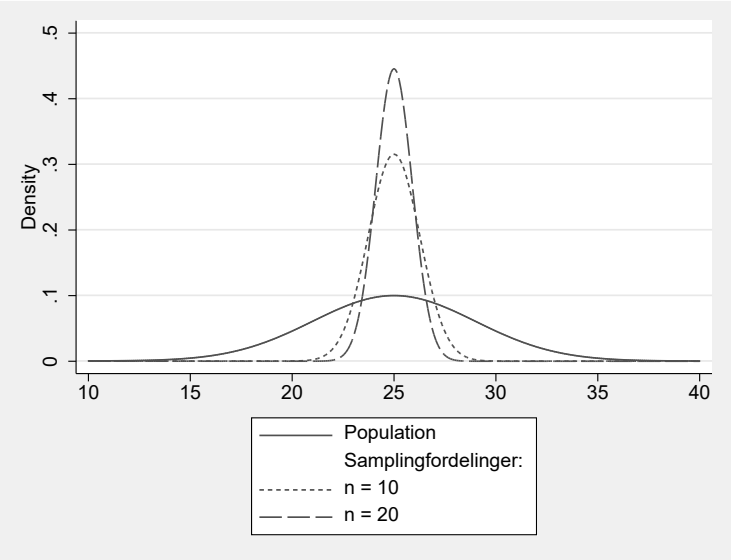
vil tilnærme sig en normalfordeling, jo større stikprøver der udtrækkes, uanset værdiernes fordeling i populationen, hedder *Den centrale grænseværdisætning*.[2]

Det er jo selvfølgelig et bemærkelsesværdigt fund, at stikprøve*gennemsnit* fordeler sig meget lig en normalfordeling, men stadigvæk: hvorfor er det så vigtigt for den teoretiske statistik, især taget i betragtning, at man sjældent tager mere en én stikprøve i en undersøgelse? Det er vigtigt, fordi *viden om*, at samplingfordelingen af gennemsnit er tilnærmelsesvis normalfordelt, gør os i stand til at konkludere angående størrelsen af et gennemsnit i en *population* på baggrund af en enkelt stikprøve. Man har blot brug for en beregning af variablens gennemsnit og standardafvigelse i den stikprøve, man har trukket, for eksempelvis at kunne konkludere, at gennemsnittet i populationen ligger inden for et ganske bestemt interval med en bestemt statistisk sikkerhed. Hvis man fx i en simpelt tilfældigt udtrukket stikprøve med 1.000 observationer har beregnet et gennemsnit og en standardafvigelse på henholdsvis 25 og 4, så kan man pba. af den teoretiske viden om samplingfordelinger samt nogle simple statistiske beregninger konkludere, at gennemsnittet i *populationen* med 95 pct. sikkerhed ligger mellem 24,75 og 25,25. Det er et eksempel på praktisk udnyttelse af teoretisk statistik, og det er langt mere tilfredsstillende end blot helt deskriptivt at sige, at vi har beregnet et gennemsnit på 25 i en tilfældig udtrukket stikprøve. Hvis vi blot gør det sidste, vil folk jo spørge, "jamen hvad så med populationen, som vi er interesserede i at vide noget om?" Læg også mærke til, at selv om der er en standardafvigelse (nogle gange også kaldet en *typisk afvigelse*) på 4 i eksemplet, så kan man i konklusionen om populationens gennemsnit med 95 pct. sikkerhed bestemme gennemsnittet inden for et meget snævert interval, nemlig på så lidt som ½, fra 0,25 under gennemsnittet i stikprøven til 0,25 over gennemsnittet i stikprøven.

Samplingfordelingen af gennemsnit er altså langt smallere end selve populations-fordelingen. Det fremgår også af Figur 2. Her vises en tænkt, normalfordelt population samt to forskellige samplingfordelinger af gennemsnit, hvor der i den ene trækkes stikprøver på 10 observationer i hver, og hvor der i den anden trækkes stikprøver med 20 i hver. Selv ved så små stikprøver kan man altså være temmelig sikker på at få et gennemsnit i sin stikprøve, der ikke ligger langt fra populationens gennemsnit. At trække en enkelt observation fra populationen med værdien 20 eller lavere ville ikke være spor underligt, men at trække en stikprøve på 10 observationer med et *gennemsnit* på 20 eller derunder vil have en nærmest forsvindende lille sandsynlighed.

[2] Flere steder på Internettet kan der findes gode, lærerige simuleringer af disse mekanismer vedrørende samplingfordelinger, fx på http://onlinestatbook.com/stat_sim/sampling_dist/

Figur 2 Population og samplingfordelinger



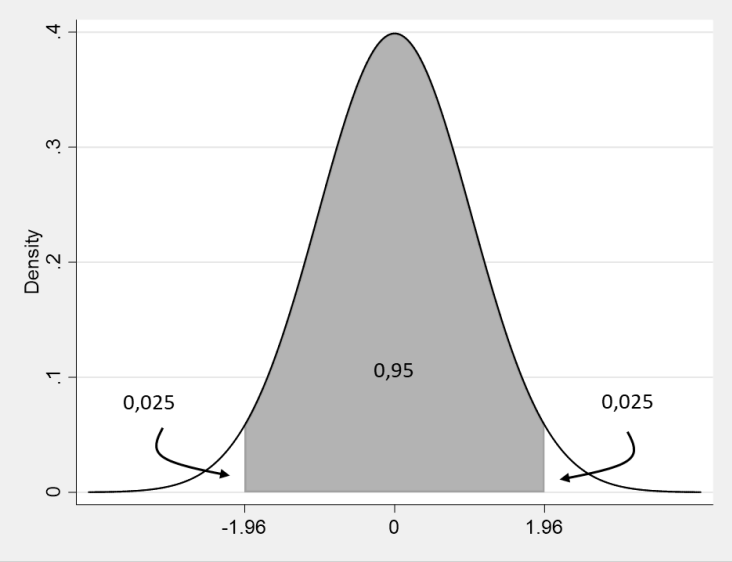
2. Sikkerhedsintervaller

Det skulle nu gerne være tydeligt, at den teoretiske statistik er smart, fordi man ved hjælp af simple beregninger på konkrete stikprøvedata samt teoretisk viden om samplingfordelinger kan konkludere ret præcist om forhold i populationen med en stor statistisk sikkerhed. Men i forrige afsnit blev der sprunget lidt let hen over nogle væsentlige detaljer. For det første spørgsmålet om, hvorfor det egentlig virker, og for det andet spørgsmålet om, hvordan man mere præcist gør for at nå frem til konklusioner om forhold i populationen. De spørgsmål vil jeg se nøjere på herunder.

Som nævnt ovenfor, er samplingfordelinger af gennemsnit tilnærmelsesvis normalfordelte ved store stikprøver. Normalfordelinger kan have større eller mindre spredning, dvs. større eller mindre standardafvigelse, hvilket også fremgår af Figur 2, men for alle normalfordelinger gælder nogle for den teoretiske statistik vigtige egenskaber. Det forholder sig nemlig sådan, at hvis man bevæger sig ud fra gennemsnittet med et bestemt antal standardafvigelser i hver sin retning, så kan man meget nemt regne ud, hvor stor en andel af observationerne der ligger herimellem. Hvis en normalfordeling fx har en standardafvigelse på 0,25, og man går to standardafvigelser til hver side herfra, dvs. 0,5 lavere end gennemsnittet og 0,5 højere end gennemsnittet, så vil en andel på cirka 0,95, altså cirka 95 pct. af observationerne, ligge i det interval. Mere præcist skal man gå 1,96 standardafvigelser på hver side af gennemsnittet for at inkludere en andel på 0,95. Det fremgår af Figur 3, og igen vises en normalfordeling i standardiseret form, hvor skalaen måler antal standardafvigelser fra gennemsnittet, dvs. at 1 på x-aksen er 1 standardafvigelse højere end gennemsnittet på 0, mens 2 er 2 standardafvigelser højere end gennemsnittet på 0 osv. Da normalfordelinger er symmetriske, vil der være en andel på 0,025 i hver hale af fordelingen,

hvis man går 1,96 ud fra gennemsnittet i hver retning, sådan som det også er vist i figuren. Og så er det vigtigt her at få med, at man nemt kan bestemme et hvilket som helst interval i en normalfordeling. Hvis man fx går 2,58 standardafvigelser på hver side af gennemsnittet, inkluderes ret præcist en andel på 0,99, og går man tre standardafvigelser på hver side af gennemsnittet, inkluderes stort set det hele, men uanset hvor langt man går ud, inkluderer man ikke 100 pct., for kurven kommer aldrig helt ned til aksen. Arealerne under kurven beregnes ved integralregning, men heldigvis er det ikke noget, man i den *praktiske* statistik vil beskæftige sig med. Før i tiden brugte man som oftest allerede beregnede tal fra tabeller, der fx viser arealer i højre hale for tallene fra 0,00 til 3,00, og ud fra de oplysninger kunne man så beregne et hvilket som helst interval med ret stor præcision. I dag er det naturligvis mere normalt at få et statistikprogram til at udføre beregningerne.

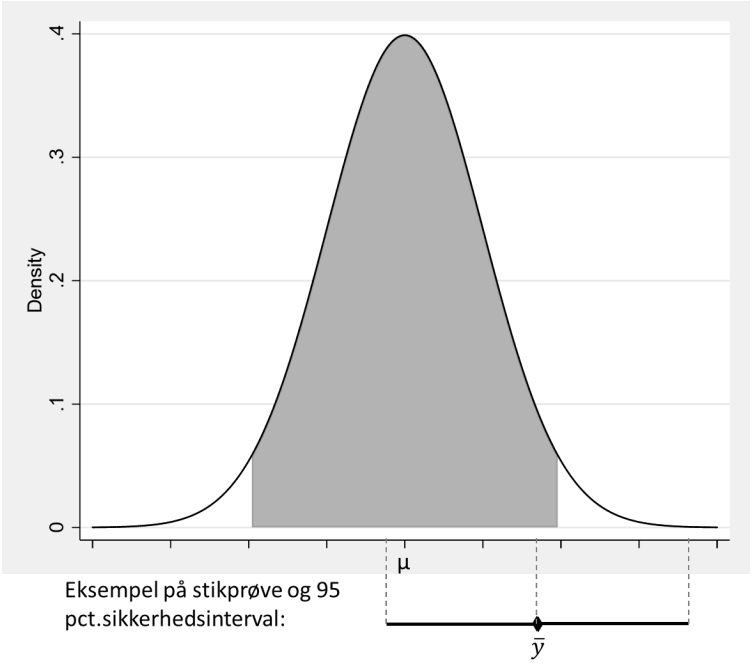
Figur 3 Andele under normalfordeling



Jeg er nu ved et vigtigt trin af argumentationskæden for, hvorfor den teoretiske statistik er smart ift. at kunne konkludere om forhold i populationen på baggrund af en stikprøve. Forestil dig, at Figur 3 viser en standardiseret samplingfordeling, fx en teoretisk fordeling af uendeligt mange stikprøvegennemsnit. Jeg skriver her en "teoretisk" fordeling, fordi man jo kun trækker en enkelt stikprøve, eller i hvert fald ikke trækker uendeligt mange. En standardafvigelse i sådan en samplingfordeling kaldes en standard*fejl*, fordi den typiske afvigelse i en samplingfordeling netop er udtryk for den typiske fejl, som man begår ift. at ramme plet i det sande gennemsnit i populationen. Værdien 0 vil her være "plet", altså udtryk for ingen afvigelse fra populationsgennemsnittet. Og fordi samplingfordelingen ligner en normalfordeling, vil 95 pct. af de uendeligt mange stikprøver have et gennemsnit, der ligger mellem -1,96 standardfejl og +1,96 standardfejl fra det sande populationsgennemsnit.

Man kan vende ovennævnte bestemmelse omvendt og direkte udlede, at en enkelt tilfældigt udtrukket stikprøve har en sandsynlighed på 0,95 for at have et gennemsnit på variablen, der ligger inden for ± 1,96 standardfejl fra populationsgennemsnittet. Det er skitseret i Figur 4 med et tænkt eksempel på en udtrukket stikprøves gennemsnit. Det græske bogstav μ (my) står for det sande populationsgennemsnit, mens mærkerne på aksen viser standardfejl. Tegnet $\bar{y}$ står for stikprøvegennemsnittet. Det fremgår tydeligt, at *hvis* gennemsnittet i den udtrukne stikprøve ligger inden for intervallet ±1,96 standardfejl fra gennemsnittet, så vil man med 100 pct. sikkerhed "fange" populationsgennemsnittet ved at lægge et interval omkring *stikprøvens* gennemsnit, hvor man går 1,96 standardfejl på hver side heraf. Derfor gælder, at man med sådan et interval omkring stikprøvegennemsnittet er 95 pct. sikker på at "fange" det sande populationsgennemsnit.

Figur 4 Skitseret samplingfordeling samt 95 pct. sikkerhedsinterval omkring stikprøvegennemsnit



Sådan et interval kaldes derfor også et 95 pct. sikkerhedsinterval, hvis formel ser ud som følger, hvor $\sigma_{\bar{y}}$ står for standardfejl i fordelingen af stikprøvegennemsnit $\bar{y}$:

$$\bar{y} \mp 1{,}96 \times \sigma_{\bar{y}} \qquad \text{(FORMEL 1)}$$

Husk på, at der er en sandsynlighed på 0,05 for, at den stikprøve, man udtrækker, havner ude i en af halerne i samplingfordelingen, hvor et sikkerhedsintervallet altså *ikke* vil "fange" det sande populationsgennemsnit. Hvis man vil være mere sikker på at fange populationsgennemsnittet, kan man gå længere ud på hver side. Jævnfør ovenfor kan man fx gå 2,58 standardfejl ud på hver side af stikprøvegennemsnittet,

sådan at man er 99 pct. sikker på at fange populationsgennemsnittet. Den ekstra sikkerhed er dog på bekostning af præcision, og det er op til forskeren at bestemme, hvilken kombination af præcision og sikkerhed der vælges. Inden for samfundsvidenskaberne vælges dog som oftest enten 95 eller 99 pct. sikkerhedsintervaller. Hvis man skriver formlen for sikkerhedsinterval uden angivelse af valgt sikkerhedsniveau, er det kutyme at skrive z for antal standardfejl:

$$\bar{y} \mp z(\sigma_{\bar{y}}) \qquad \text{(FORMEL 2)}$$

Som nævnt i ovenstående, var det især tidligere almindeligt at slå op i en tabel for at finde z-værdier tilhørende en bestemt sandsynlighed i højre hale, eller omvendt at finde den sandsynlighed, der hører til en bestemt z-værdi. En sådan tabel kan ses i artiklens Bilag 2. I dag er det langt mere almindeligt at foretage den slags ved hjælp af et statistikprogram, en lommeregner eller en hjemmeside på Nettet. I Boks 1 vises, hvordan man kan komme fra det ene til det andet i Statistikprogrammet Stata.

BOKS 1

FRA Z TIL P ELLER OMVENDT I STATA

Fra sandsynlighed i højre hale til z-værdi:

Hvis man fx vil finde en z-værdi til en p-værdi for højre hale på 0,05:

```
. display invnormal(1-0.05)
1.6448536
```

Fra z-værdi til sandsynlighed i højre hale:

Hvis man fx vil finde p-værdi tilhørende en z-værdi på 1,5:

```
. display 1-normal(1.96)
.0249979
```

Hvis dobbeltsidet p-værdi (begge haler):

```
. display 2*(1-normal(1.96))
.04999579
```

Et eksempel med større præcision i z-værdi, jævnfør Eksempel 2 i Afsnit 3:

```
. display 2*(1-normal(2.3918))
.01676598
```

Den grundlæggende introduktion til mekanismerne i den teoretiske statistik via estimering af sikkerhedsinterval omkring et gennemsnit på en intervalskaleret variabel skulle nu være på plads, men der mangler stadigvæk noget, førend man kan foretage beregningerne. Vi mangler størrelsen på standardfejlen, dvs. standardafvigelsen på den

teoretiske samplingfordeling af stikprøvegennemsnit. Hvis stikprøven er simpelt tilfældigt udtrukket fra hele populationen, er formlen for denne[3]:

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} \qquad \text{(FORMEL 3)}$$

Heraf ses det, at des større variation på variablen (hvor σ står for standardafvigelsen), des større spredning er der i samplingfordelingen. Det kan man som forsker ikke gøre så meget ved. Variablen er, som den nu er. Det fremgår endvidere af formlen, at des større stikprøver, der udtrækkes, des mindre spredning i samplingfordelingen. Dvs. at man kan mindske den statistiske usikkerhed ved at udtrække en større stikprøve, hvilket ikke er så overraskende.

Man løber imidlertid ind i et problem, når man skal udregne standardfejlen, for man kender sjældent *σ*, dvs. standardafvigelsen på variablen i hele populationen. Statistikere har imidlertid fundet ud af, at hvis blot man udtrækker en stikprøve af en vis størrelse, så sker der kun en lille fejl ved at estimere standardafvigelsen i populationen med standardafvigelsen i stikprøven, dvs. udskifte *σ* med *s* i ovenstående formel således:

$$se = \frac{s}{\sqrt{n}} \qquad \text{(FORMEL 4)}$$

Læg mærke til at der nu på venstresiden af ligningen står *se*. Det er for at angive, at det ikke er den sande standardfejl, der beregnes, men derimod en *estimeret* standardfejl[4]. Med estimering af standardfejlen ser formlen for sikkerhedsinterval sådan her ud:

$$\bar{y} \mp z(se) \qquad \text{(FORMEL 5)}$$

I praksis vil fejlen ved denne udskiftning være acceptabelt lille, hvis den udtrukne stikprøve har en vis størrelse, ofte nævnes en minimumsgrænse på 30 observationer. I afsnit 4 skal det dog gennemgås, hvordan man kan tage højde for den fejl, der trods alt

[3] Af pædagogiske hensyn, og da artiklen skal ses som introducerende til grundmekanismerne i inferens fra stikprøve til population, vil der igennem artiklen forudsættes simpelt tilfældigt udtræk, hvor hver enhed fra populationen har den samme sandsynlighed for at blive udtrukket til stikprøven. I den virkelige verdens stikprøver ifm. fx surveyundersøgelser vil dette imidlertid sjældent være tilfældet, og i så fald bliver estimeringen mere kompliceret. Som oftest vil der ganske vist være tale om en eller anden form for sandsynlighedsudvælgelse, men det vil typisk være i form af en klyngeudvælgelse, en stratifikationsudvælgelse eller en kombination heraf. Ydermere vil der meget ofte vil være et skævt frafald fra bruttostikprøve til nettostikprøve, fordi man ikke altid kan finde frem til de personer, man udvælger, eller fordi nogle personer ikke har lyst til at deltage, samtidig med at dette frafald ikke er tilfældigt fordelt blandt de udvalgte til bruttostikprøven. Af disse årsager vil man ofte vægte data tilbage til populationsrepræsentativitet (så godt man kan), og der vil i beregningen af standardfejl blive korrigeret for den som oftest forøgede usikkerhed i estimeringen. I statistikprogrammet Stata foretages disse vægtninger og justeringer af signifikansestimeringen via svy-kommandoer.

[4] Jeg benytter samme forkortelse af den estimerede standardfejl som i:
Agresti, Alan (2018). Statistical Methods for the Social Sciences, 5th edition, Pearson.

vil være til stede, uanset om den er lille, men først skal der herunder gennemgås et konkret eksempel via simpel metode ved store stikprøver.

EKSEMPEL 1: 95 PCT. SIKKERHEDSINTERVAL FOR INTERVALSKALEREDE VARIABLER:

Jeg vil her vende tilbage til det fiktive eksempel, som jeg indledte kapitlet med. Her noterede jeg mig et stikprøvegennemsnit på 25 samt en standardafvigelse på 4, og stikprøven var et simpelt tilfældigt udtræk på 1.000 observationer. Førend jeg kan estimere et 95 pct. sikkerhedsinterval, skal standardfejlen estimeres (fra formel 4):

$$se = \frac{s}{\sqrt{n}} = \frac{4}{\sqrt{1000}} = 0{,}1265$$

Jeg viser her et resultat afrundet til fire decimaler. Normalt vil man ikke afrunde mellemresultater, men i stedet gemme disse på sin lommeregner, eller hvad man nu bruger. Den typiske fejl, der sker ved estimering af populationsgennemsnittet ud fra stikprøvegennemsnittet i den givne situation, er altså cirka 0,1265. Hvis man nu går 1,96 gange denne størrelse på hver side af stikprøvegennemsnittet, er man 95 pct. sikker på, at man vil fange det sande populationsgennemsnit. Det ser sådan her ud (fra formel 5):

$$25 \mp 1{,}96 \times 0{,}1265$$

Hvis man udregner multiplikationen, får man det, der kaldes *fejlmargin*, hvorfor sikkerhedsintervallet også siges at være lig med stikprøvegennemsnittet plus/minus en fejlmargin:

$$25 \mp 0{,}2479$$

Intervallet kan selvfølgelig også skrives på følgende facon, som nogle gange kan forekomme lidt lettere at skrive en kort opsummerende tekst til:

$$[24{,}75;\ 25{,}25]$$

Med 95 pct. sikkerhed ligger gennemsnittet i populationen altså mellem 24,75 og 25,25 (afrundet til to decimaler efter kommaet).

3. Hypotesetest

Jeg vil nu gennemgå principperne for statistisk hypotesetest, og ligesom ved forrige afsnit om sikkerhedsintervaller bliver gennemgangen eksemplificeret ved en simpel én af slagsen, nemlig en test for størrelsen af et gennemsnit på en variabel. Den type af test ses ganske vist forholdsvis sjældent inden for samfundsvidenskabelig forskning. Til gengæld er det forholdsvis let fordøjeligt, og det er nøjagtigt de samme

hovedprincipper, der bruges ved langt de fleste andre typer af statistiske hypotesetest, fx for korrelationskoefficienter, koefficienter ved både lineær og logistisk regression og forskel i gennemsnit mellem to grupper. Så viden om principperne for hypotesetest er nærmest uomgængeligt, hvis man vil beskæftige sig med analyse af kvantitative data, nøjagtigt som det gælder for sikkerhedsintervaller.[5]

Statistisk hypotesetest går ud på, at der til en start opstilles en hypotese angående et forhold i den population, som ens stikprøve er udtrukket fra. Hypotesen kaldes normalt for en *nul-hypotese* eller blot H_0, og den kan fx være en påstand om, at gennemsnittet i populationen på en variabel har en bestemt værdi, eksempelvis at danskere i gennemsnit ser "gammeldags" flow-TV i 120 minutter dagligt, eller at danskere på 18 år og derover har et gennemsnit på 5 på en selvvurderet, politisk venstre/højre-skala fra 0 til 10. Hvis vi griber fat i sidstnævnte eksempel, så går selve testen nu ud på at undersøge, hvorvidt dataene fra stikprøven taler så meget imod denne hypotese, at man med tilstrækkelig stor statistisk sikkerhed kan afvise den. Det har jo klart noget at gøre med, hvor langt stikprøvens gennemsnit ligger fra nulhypotesens værdi, hvor stor spredning der er i stikprøvens data, og hvor stor en stikprøve man har udtrukket. Jo længere væk stikprøvens gennemsnit ligger fra nulhypotesen, des mere taler det imod nulhypotesen, og jo mindre spredning der samtidig hermed er i stikprøvens data, og jo større stikprøve der er udtrukket, des mere sikkert kan man afvise nulhypotesen. Og *hvis* man på et sikkert statistisk grundlag kan afvise nulhypotesen, i det her tilfælde at den gennemsnitlige venstre/højre-placering er lig med 5, så betyder det jo omvendt, at man kan acceptere en alternativ hypotese, ofte blot kaldet H_A, der siger, at gennemsnittet i populationen er forskelligt fra 5.

Nu er det måske ikke så fantastisk at kunne udelukke et enkelt punkt på talrækken som populationsgennemsnit, men heldigvis kan vi faktisk også mere end det. Hvis fx at stikprøvegennemsnittet viser sig at være *højere* end 5, dvs. at der er et gennemsnit til højre for den politiske midte, og at dataene taler tilstrækkeligt statistisk sikkert imod nulhypotesen om et populationsgennemsnit på 5, så kan vi samtidig afvise, at populationsgennemsnittet er lavere end 5, dvs. til venstre for midten. Det kan vi, fordi alle tænkelige nulhypoteser med gennemsnit på under 5 vil betyde, at stikprøvens data blot taler endnu mere imod nulhypotesen, fordi afstanden mellem nulhypotesen og stikprøvegennemsnittet så bliver endnu større. Så selvom man i en hypotesetest formelt set og umiddelbart blot afviser et enkelt tal, fx 5, så vil man i det pågældende eksempel sige, at der implicit testes en nulhypotese, der lyder at gennemsnittet i populationen er 5 eller derunder. Det skal i øvrigt lige bemærkes, at man selvfølgelig kan få et stikprøvegennemsnit, der er lavere end 5, og hvis nulhypotesen i den situation afvises, vil det implicit være en nulhypotese, der siger, at populationsgennemsnittet er

[5] Dette også selvom der er en stigende erkendelse af begrænsningerne i hypotesetest samt efterhånden en udbredt diskussion om statistisk signifikans i det hele taget. Se fx artiklen "The ASA's Statement on *p*-Values: Context, Process, and Purpose" af Ronald L. Wasserstein & Nicole A. Lazar i *The American Statistician*, Vol. 70, 2016, Issue 2.

5 eller derover, der reelt afvises. Endelig kan vi jo komme ud for, at dataene i stikprøven ikke taler stærkt nok imod nulhypotesen til at man kan afvise den. Mange vil så sige, at man accepterer nulhypotesen, men det ville ikke være korrekt. Dataene i stikprøven taler blot ikke stærkt nok imod den, til at man kan afvise den, men mere herom nedenfor.

Hvor man i estimationen af et sikkerhedsinterval tager udgangspunkt i stikprøvens data, går man i udførelsen af en hypotesetest i stedet ud fra nulhypotesen. Man antager, at denne er korrekt, og hvis den er det, dvs. hvis gennemsnittet i populationen er, som nulhypotesen siger, så vil samplingfordelingen, altså den hypotetiske/teoretiske fordeling af mulige stikprøvegennemsnit, ligge omkring nulhypotesens værdi. Hvis fx nulhypotesen i ovennævnte eksempel er korrekt, så vil samplingfordelingen ligge omkring 5. Samplingfordelingen ville altså have et gennemsnit på 5, og den ville være tilnærmelsesvis normalfordelt. Man ved ikke præcist, hvilken standardafvigelse fordelingen vil have, men man kan ligesom ved sikkerhedsintervaller estimere denne ved udskiftning af σ med s i formlen for standardfejl:

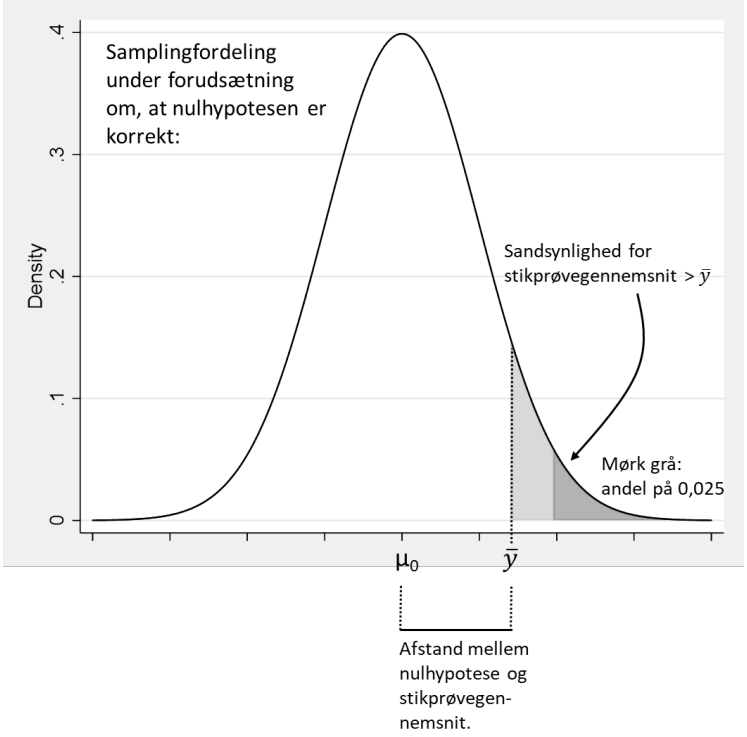
$$se = \frac{s}{\sqrt{n}} \qquad \text{(FORMEL 4 GENTAGET)}$$

Hvis vi ud fra stikprøvedataene udregner ovenstående estimerede standardfejl, er næste skridt at beregne, hvor stor afstanden er mellem nulhypotesens påstand om populationsgennemsnit, μ_0, og stikprøvens gennemsnit, $\bar{y}$. Denne afstand divideres nu med den estimerede standardfejl, sådan at vi får afstanden udtrykt i antal standardfejl, *z*:

$$z = \frac{\bar{y} - \mu_0}{se} \qquad \text{(FORMEL 6)}$$

Læg mærke til, at hvor man ifm. sikkerhedsintervaller *beslutter* sig for en bestemt statistisk sikkerhed, fx 95 pct., og dernæst vælger den tilhørende *Z*-værdi, så *beregner* man ifm. hypotesetest *z*-værdien via formel 6. I Figur 5 anskueliggøres principperne ved hypotesetest, og her fremgår også *z*, nemlig afstand mellem μ_0 og $\bar{y}$, udtrykt i antal standardfejl. Da samplingfordelingen er tilnærmelsesvis normalfordelt, kan vi ved hjælp af *z*-værdien beregne arealet af kurven, der ligger længere ude end *z*. Her udnytter man igen kendte egenskaber ved normalfordelinger, hvor alle arealer over intervaller på den horisontale akse kan beregnes. Jævnfør ovenfor om sikkerhedsintervaller forholder det sig fx sådan, at går man 1,96 standardfejl over gennemsnittet, findes en sandsynlighed på 0,025 i arealet i højre hale. Spørgsmålet er så, hvor stor en sandsynlighed, der hører til den beregnede *z*-værdi. Det benytter man tabeller, lommeregner eller statistikprogrammer til at beregne, og i eksemplet, der er afbilledet i Figur 5, er *z*-værdien lig med 1,4 (altså 1,4 standardfejl højere end nulhypotesens bud på gennemsnit i populationen), og det tilhørende areal kan beregnes til 0,081. Jeg vil nedenfor gennemgå et konkret eksempel lidt mere udførligt. Lige nu drejer det sig først og fremmest om principperne

Figur 5 Principper ved hypotesetest



Men hvad skal man så konkludere på baggrund af det statistiske resultat? Ja, umiddelbart kan man konkludere, at der er en sandsynlighed på 0,081 for at trække en stikprøve med mindst så højt et gennemsnit som $\bar{y}$ under forudsætning om, at nulhypotesen er korrekt (det sidste, angående forudsætningen, er en vigtig tilføjelse, som man ofte ser glemt). Nu går den statistiske hypotesetest i første omgang ud på at vurdere, om man vil forkaste nulhypotesen. Først efter en eventuel forkastelse af den, kan man godtage den alternative. Det kaldes på engelsk for *proof by contradiction*. Og det afgørende for valget om forkastelse af nulhypotesen er, hvor lidt sandsynligt det er at trække en stikprøve mindst lige ekstrem som det fundne, hvis nulhypotesen er korrekt. Man bør derfor vurdere sandsynligheden for at trække en stikprøve med et gennemsnit på $\bar{y}$ eller højere *eller* på $-\bar{y}$ eller lavere, givet at nulhypotesen er korrekt. Så man skal finde p-værdien i venstre hale også, men da en normalfordeling er symmetrisk, er det blot at gange p-værdien i højre hale med 2. Under antagelse af at nulhypotesen er korrekt, betyder det, at sandsynligheden for at trække en stikprøve med et gennemsnit, der afviger mindst lige så meget fra nulhypotesen, som det vi har fundet, er 0,162 ($2 \times 0{,}081$), eller 16,2 pct. Det er selvfølgelig en subjektiv vurdering, hvor meget der skal til, førend man synes, at stikprøvens data taler sikkert nok imod nulhypotesen, til at man kan afvise denne, men det er kutyme at bruge nogle ret små og runde tal som en slags grænseværdier, fx 0,05 eller 0,01, svarende til henholdsvis 5 og 1 pct. En sådan grænse angives også ofte med det græske bogstav α (alfa), mens sandsynlighederne i fordelingens haler enkeltvis kaldes for $\alpha/2$. Alfa-niveauet kaldes også for *signifikansniveauet*. I eksemplet her fik jer en p-værdi på 0,162, og det er under alle

omstændigheder for højt til at man vil afvise nulhypotesen. Man kan jo ikke sige, at nulhypotesens påstand om et gennemsnit på 5 virker urealistisk, hvis der i givet fald er over 16 pct. sandsynlighed for at ramme mindst så langt fra 5, som man har gjort med sin stikprøve. Man siger i den situation også, at testen er statistisk insignifikant.

Den slags test, hvor man medtager begge haler i sin endelige p-værdi, kaldes for dobbeltsidede hypotesetest. Der findes også enkeltsidede test, hvor den alternative hypotese kun angår den ene eller den anden retning fra nulhypotesen, men de er ikke gængse, i hvert fald ikke inden for samfundsvidenskaben, og det er meget omdiskuteret, om og i givet fald i hvilke situationer, enkeltsidede test kan retfærdiggøres. Men én ting ligger fast: man må ikke bestemme sig for enkeltsidet test og derefter lade stikprøvedata afgøre, hvilken side man tester mod. Det vil blive betragtet som på uredelig vis at halvere den reelle p-værdi. Hvis man beslutter sig for enkeltsidet test, vil begrundelsen være, at man kun kan forestille sig et sandt populationsgennemsnit i den ene retning fra nulhypotesen. Selv da vil der ofte kunne argumenteres for brug af dobbeltsidet test, men ikke mere om det her. Så i den resterende tekst vil jeg alene omtale dobbeltsidede test.

Inden jeg går videre med endnu et eksempel på hypotesetest, skal de trin, som man altid (mere eller mindre eksplicit) går igennem, ridses op herunder[6]:

1. Er forudsætningerne for at udføre netop denne test til stede? Er variablen fx på det korrekte måleniveau, og er stikprøven trukket simpelt tilfældigt? I situationen med test for gennemsnit skal man fx sikre sig, at variablen er intervalskaleret eller ratioskaleret, så det giver mening at beregne gennemsnit på den.

2. Der opstilles nulhypotese og alternativ hypotese.

3. Der beregnes *teststatistik*. Teststatistikken er et komprimeret mål for, hvor meget stikprøvedataene afviger fra nulhypotesen. I ovenstående har jeg beskrevet teststatistikken z, der er et mål for antal standardfejl, stikprøvegennemsnittet falder fra nulhypotesens påstand, men der findes andre teststatistikker til andre typer af hypotesetest, og i afsnit 4 gennemgås teststatistikken t.

4. Der beregnes en til teststatistikken tilhørende p-værdi. Jo mindre p-værdi, des større er tilskyndelsen til at afvise nulhypotesen og godtage den alternative hypotese.

5. Fortolkning af p-værdien og evt. en formel beslutning om afvisning af nulhypotese pba. en bestemt α-værdi, fx 0,05.

[6] Frit efter Agresti (2018), p154.

Det er vigtigt at gennemgå alle trinnene, når man udfører en hypotesetest, men det behøves ikke at gøres meget udspecificeret punktvis. Det må gerne gøres lidt flydende i en samlet tekst, sådan som jeg fx gennemgår det følgende eksempel.

EKSEMPEL 2: HYPOTESETEST FOR INTERVALSKALEREDE VARIABLER:

I ovenstående gennemgang af principperne for hypotesetest har jeg stort set samtidig gennemgået et eksempel. Der blev imidlertid sprunget lidt let hen over nogle udregninger, hvorfor jeg nedenfor mere slavisk vil gennemgå endnu et eksempel. Igen drejer det sig om politisk selvplacering på en venstre/højre-skala fra 0 til 10, altså med en midterværdi på 5. På stikprøvedata fra den danske del af European Social Survey 2017 finder jeg en gennemsnitlig selvplacering på 5,3266 for de unge på 16-29 år med en standardafvigelse på 2,3533. Spørgsmålet er, om der er tilstrækkelig stor statistisk sikkerhed for, at de unge gennemsnitligt ligger til højre for midten[7]. Hvis jeg kan afvise en nulhypotese, der lyder, at populationsgennemsnittet for de 16-29 årige er lig med 5, kan jeg godtage en alternativ hypotese om, at populationsgennemsnittet forskellig fra 5, samt reelt også at det må være større end 5, jævnfør den indledende diskussion i indeværende afsnit. Med andre ord vil man i givet fald kunne konkludere, at populationsgennemsnittet på politisk selvplacering ligger til højre for midten i populationen. Formelt kan hypoteserne i den dobbeltsidede hypotesetest opstilles således:

$$H_0: \mu = 5$$

$$H_a: \mu \neq 5$$

Stikprøvestørrelsen, dvs. *n* for denne delpopulation af 16+ årige danskere er på 297 respondenter, og jeg har nu de oplysninger, der skal til for at foretage hypotesetesten. Forudsætningerne for at gennemføre testen er opfyldt. Stikprøven er udtrukket simpelt tilfældigt, og variablen er intervalskaleret. Det sidste kan muligvis diskuteres. Nogle vil måske mene, at den kun er ordinalskaleret, men da respondenterne kun ser skalaen med cifrene 0 til 10 ud over en tekst ved de to poler for mest venstreorienteret og mest højreorienteret, vil jeg betragte den som intervalskaleret.

Efter således at have diskuteret forudsætningerne, er det næste skridt at estimere standardfejlen i den hypotetiske samplingfordeling, gældende for nulhypotesen om et gennemsnit på 5 (fra formel 4):

[7] Om de unge ligger til højre for midten kan måles på forskellig vis. Her er det den gennemsnitlige placering, der testes på, men man kunne jo også sammenligne andele under og over 5, hvilket ikke nødvendigvis vil give sammen konklusion.

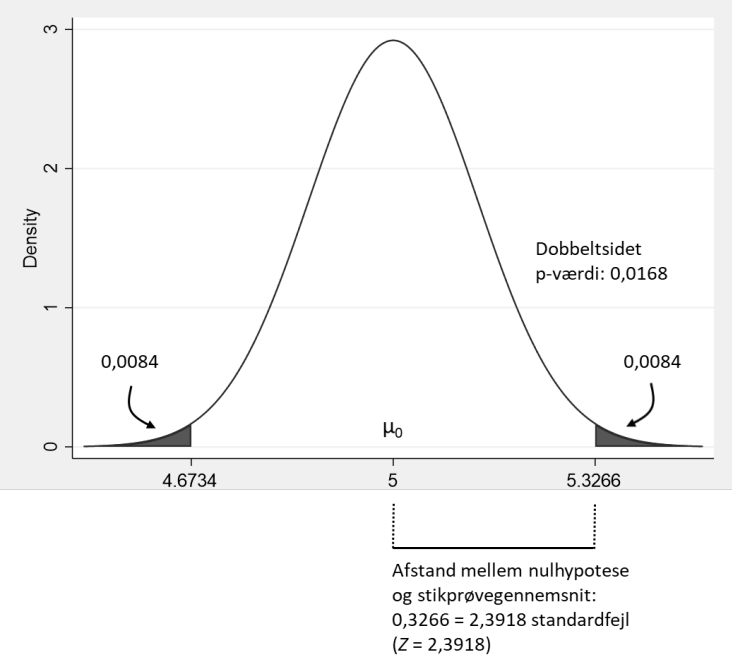
$$se = \frac{s}{\sqrt{n}} = \frac{2{,}3533}{\sqrt{297}} = 0{,}1366$$

Herefter måles afstanden mellem stikprøvegennemsnittet og nulhypotesens påstand om gennemsnit, og afstanden konverteres til antal standardfejl, dvs. teststatistikken beregnes (fra formel 6)[8]:

$$z = \frac{\bar{y} - \mu_0}{se} = \frac{5{,}3266 - 5}{0{,}1366} = 2{,}3918$$

Denne afstand fremgår også af Figur 6, som viser samplingfordelingen under forudsætning af korrekt nulhypotese. Endelig skal den til z-værdien hørende sandsynlighed findes. Husk at højre hale til $z = 1{,}96$ er lig med 0,025, og vi er længere ude på fordelingen her, nemlig $z = 2{,}3918$. Den tilhørende enkeltsidede p-værdi (dvs. højre hale) er lig med 0,0084, så den dobbeltsidede p-værdi er lig med 0,0168 ($2 \times 0{,}0084$). Nederst i Boks 1 er vist, hvordan disse tal kan findes via Stata. Med en præcision på to decimaler efter kommaet på z-værdien kan det også slås op i standard-normalfordelingstabellen i artiklens Bilag 2.

Figur 6 Dobbeltsidet hypotesetest



Så mangler jeg blot at konkludere på resultatet. I første omgang er konklusionen følgende: hvis nulhypotesen er korrekt, er der en sandsynlighed på 0,0168 for at få et stikprøvegennemsnit, der er mindst lige så afvigende, som det, jeg har fundet. Spørgsmålet er så, om den sandsynlighed er så lille, at jeg vælger at afvise nulhypotesen.

[8] Resultatet er fra udregning med ikke afrundet standardfejl fra den forrige udregning.

Som nævnt ovenfor, benytter man ofte små runde tal som 0,05 og 0,01 som grænseværdier. Her har vi en p-værdi på 0,0168, så den er klart under 0,05, men ikke under 0,01. I sådan en situation vil man ofte afvise nulhypotesen og i sin tekst skrive noget i retning af, at resultater er statistisk signifikant på 0,05-niveau.

Hvis man får en p-værdi på fx 0,06 kan man dog sagtens stadigvæk bemærke det som interessant og fx skrive noget i retning af, at det er meget tæt på at være statistisk signifikant på 0,05-niveau. De præcise grænseværdier som 0,05 og 0,01 er jo valgt, fordi de er pæne runde tal, men det gør dem jo ikke magiske. Man lægger heller ikke så meget vægt på grænseværdierne i dag, som man gjorde tidligere. Dvs. at man ikke på forhånd bestemmer sig for et bestemt signifikansniveau. Og i hvert fald er det i dag kutyme at notere selve p-værdien, dvs. ikke kun angive om en test er statistisk signifikant på fx 0,05-niveau. Så kan den oplyste læser selv vurdere sikkerheden. Og egentlig behøver man ikke være særlig oplyst. Det kræver blot, at man ved, hvad en sandsynlighed er for noget. I sammenhæng hermed skal det tilføjes, at der måske nogle gange bliver lagt for meget vægt på hypotesetest ift. sikkerhedsintervaller. Som Agresti (2008) skriver, er sikkerhedsintervaller ofte mere informative. At en hypotesetest viser, at et populationsgennemsnit med stor statistisk sikkerhed fx er større end det, som nulhypotesen påstår, gør ikke umiddelbart én meget klogere på en mere præcis værdi af populationsgennemsnittet. Det er til dels rigtigt, men på den anden side oparbejder man også efterhånden en erfaring, der gør, at man pba. p-værdi, stikprøvegennemsnit og stikprøvestørrelse formår at danne sig et ret godt indtryk af sandsynlige værdier af populationsgennemsnittet, selvom det selvfølgelig ikke er med samme præcision, som et formelt sikkerhedsinterval kan gøre det. En anden ting er, at hypotesetest i situationer med mange estimerede parametre, som fx ifm. multiple regressions-modeller, kan være noget nemmere at overskue for læseren end sikkerhedsintervaller.

Principperne for hypotesetest og sikkerhedsintervaller er nu gennemgået sammen med eksempler på begge, og det skulle være tydeligt, at de to er i familie med hinanden. Standardfejlen estimeres på samme måde, og der er fuld overensstemmelse mellem et 95 pct. sikkerhedsinterval og en hypotesetest med et signifikansniveau på 0,05. Hvis at man således på baggrund af en hypotesetest, som i ovenstående eksempel, afviser nulhypotesen på 0,05-niveau, så vil et 95 pct. sikkerhedsinterval heller ikke inkludere nulhypotesens værdi. Og hvis man omvendt i en hypotesetest *ikke* formår at afvise nulhypotesen på 0,05-niveau, så *vil* nulhypotesens værdi også inkluderes i et 95 pct. sikkerhedsinterval.

4. Hypotesetest og sikkerhedsinterval ved benyttelse af *t*-fordelingen

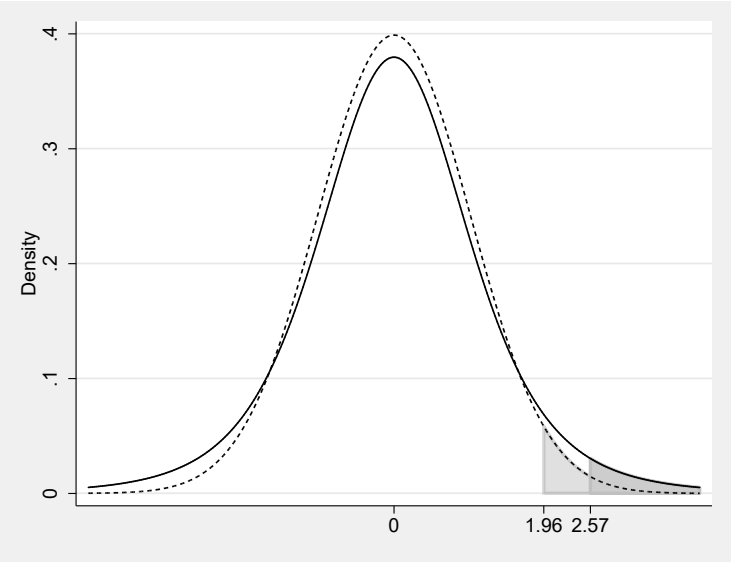
I de foregående afsnit mangler der en diskussion af en vigtig detalje for inferens fra stikprøve til population, der i nogle situationer kan give et problem. Jeg har i afsnit 1

beskrevet, dels hvordan samplingfordelingen af stikpøvegennemsnit tilnærmer sig en normalfordeling, jo større stikprøver der udtrækkes, uanset fordelingen i populationen, dels at samplingfordelingen altid vil være normalfordelt, hvis variablen i populationen er normalfordelt. Med andre ord kan vi bruge metoderne til inferens, der er beskrevet i afsnit 2 og 3 angående sikkerhedsintervaller og hypotesetest, hvis vi er sikre på, at variablen er nogenlunde normalfordelt i populationen, og/eller at stikprøven er stor. Selv i ret så små stikprøver skal der således en del afvigelse fra normalfordeling til, førend der sker betydelige fejl. Så det problem er til at overskue, og under alle omstændigheder svært at gøre noget ved.

Der er imidlertid endnu et problem, som også er nævnt i afsnit 1. Vi kan ikke direkte beregne den korrekte standardafvigelse i samplingfordelingen, dvs. standardfejlen. Vi kan kun *estimere* den ved en udskiftning af σ med *s* i formlen for standardfejlen (formel 4), med mindre at man skulle være så heldig at kende standardafvigelsen i populationen. Heldigvis er fejlen ved at udskifte σ med *s* marginal i store stikprøver. Men fejlen kan være ganske betydelig i meget små stikprøver, uanset om variablen er normalfordelt i populationen eller ej. Tilbage i det første årti af 1900-tallet fandt en britisk statistiker[9] imidlertid ud af, at man kan vurdere størrelsen af den fejl, og at man kan beregne en sandsynlighedsfordeling, der i modsætning til standardnormalfordelingen tager højde for den. Han analyserede ofte ret små stikprøver med byg i sit arbejde med brygning af øl på Guinness bryggeriet, og selvom populationen, som han trak stikprøver fra, var normalfordelt, så passede beregningerne ikke helt pga. fejlen ved estimeringen af standardfejlen. Han fandt frem til en familie af sandsynlighedsfordelinger, der afhænger af stikprøvestørrelsen, sådan at fordelingen ved små stikprøver er en del bredere end standardnormalfordelingen, og jo større stikprøve, des mere ligner fordelingen en standardnormalfordeling. Dvs. at man ved analyse på en lille stikprøve ikke blot skal over en teststatistik-værdi på 1,96 for at få et resultat, der er statistisk signifikant på 0,05-niveau. Man skal op til en større værdi. Fundet revolutionerede statistikken med små stikprøver, og statistikprogrammer benytter selv ved store stikprøver *t*-fordelinger i stedet standardnormalfordelingen ved en række forskellige statistiske metoder, som fx test for gennemsnit og koefficienter i lineær regression. Fejlen ved estimeringen af standardfejlen er der jo altid. Den bliver blot mindre og mindre, des større stikprøve man trækker. Den konkrete t-fordeling betegnes ved noget, der i statistikken kaldes for *frihedsgrader*, forkortet *df* for "degrees of freedom". Ved simpel test for gennemsnit er *df* lig med antal enheder i stikprøven minus 1, altså $n - 1$. I Figur 7 er vist dels standardnormalfordelingen, dels en t-fordeling med $df = 5$, dvs. ved stikprøver på 6 enheder.

[9] William Sealy Gosset (med pseudonymet Student).

Figur 7 $\alpha/2$ **= 0,025 for henholdsvis t-fordeling med df = 5 og standardnormalfordeling**

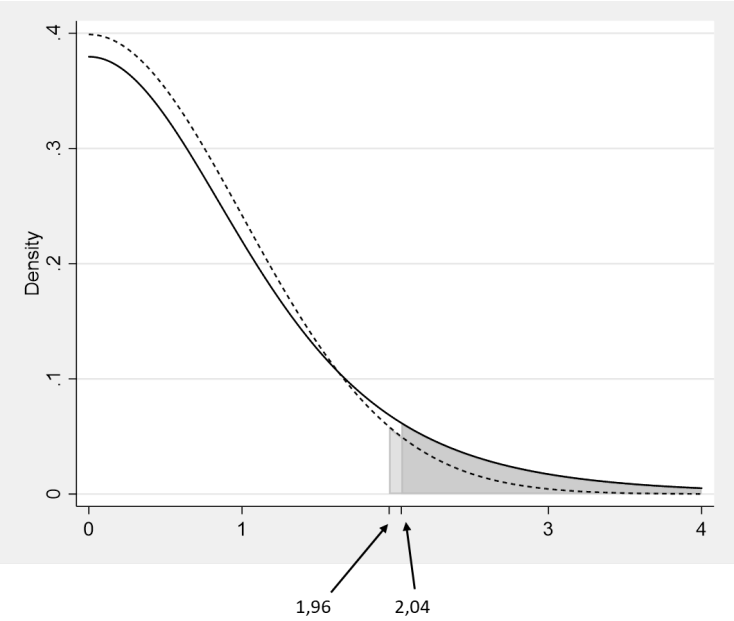


Fuldt optrukket linje: t-fordeling med df = 5
Stiplet linje: standard normalfordeling

Det fremgår af figuren, at *t*-fordelingen med $df = 5$ er en noget fladere end standardnormalfordelingen, og man skal ud på 2,57 førend der kun er 2,5 pct. tilbage i højre hale, mens det ved standardnormalfordelingen er 1,96. Som det imidlertid fremgår af Figur 8, skal man ikke op på en synderligt stor stikprøve, førend *t*-fordelingen ligner en standardnormalfordeling næsten til forveksling. Her sammenlignes standardnormalfordelingen med en t-fordeling med $df = 30$.

Ved at benytte en t-fordeling, der passer til stikprøvestørrelse tager man altså højde for den ekstra usikkerhed, der opstår, hvis man – som i det fleste tilfælde – ikke kender den sande standardafvigelse i populationen, men i stedet må estimere den ved standardafvigelsen i stikprøven. I Boks 2 vises, hvordan man i Stata kan jonglere rundt mellem *t*-værdi og p-værdi.

Figur 8 **$\alpha/2$ = 0,025 for henholdsvis t-fordeling med df = 30 og standardnormalfordeling**



Fuldt optrukket linje: t-fordeling med *df* = 30
Stiplet linje: standard normalfordeling

BOKS 2

FRA *t* TIL p ELLER OMVENDT I STATA

Fra sandsynlighed i højre hale til t-værdi:

Find *t*-værdi med df = 24 til en p-værdi for højre hale på 0,025, jævnfør Eksempel 4 nedenfor:

```
. display invttail(24,0.025)
2.0638986
```

Find *t*-værdi med df = 296 til en p-værdi for højre hale på 0,025:

```
. display invttail(296,0.025)
1.9680107
```

Fra t-værdi til sandsynlighed i højre hale:

Find p-værdi i højre hale tilhørende en *t*-værdi på 1,5 med df = 24:

```
. display ttail(24,1.5)
.07332782
```

Find dobbeltsidet p-værdi med tilhørende *t*-værdi på 2,3918, jævnfør Eksempel 3 nedenfor:

```
. display 2*ttail(296,2.3918)
.01739
```

EKSEMPEL 3: HYPOTESETEST MED GENNEMSNIT VED BENYTTELSE AF T-FORDELING

Jeg gentager testen på data fra European Social Survey med en nulhypotese om gennemsnit i populationen på politisk selvplacering blandt 16-29-årige på 5. Standardfejlen estimeres på samme måde, og den blev ovenfor beregnet til 2,3918. Teststatistikken beregnes også på samme måde, men den betragtes nu ikke som en z-værdi, men derimod som en *t*-værdi for at medregne den ekstra statistiske usikkerhed i ikke at kende populationens standardafvigelse:

$$t = \frac{\bar{y} - \mu_0}{se} = \frac{5{,}3266 - 5}{0{,}1366} = 2{,}3918$$

I det forrige eksempel, hvor jeg benyttede en z-statistik, fik jeg en enkeltsidet p-værdi på 0,0084 og en dobbeltsidet på 0,0168. I *t*-fordelingen med df=296 ligger der en større andel ude i halerne, og her får jeg i stedet en enkeltsidet p-værdi, der er en anelse større, nemlig 0,0087, og en dobbeltsidet p-værdi på 0,0174 ($2 \times 0{,}0087$). I Boks 2 nederst vises, hvordan man kan få beregnet p-værdien i Stata. Fordi det er en forholdsvis stor stikprøve, ligner tallene fra de to metoder hinanden. Kurverne ligger praktisk talt oven i hinanden, og det vil være meningsløst at forsøge at sammenligne de to kurver i en figur. Konklusionen er også den samme i begge analyser, nemlig at nulhypotesen om et populationsgennemsnit på 5 kan afvises på 0,05-niveau, men ikke på 0,01-niveau. Det er dog stadigvæk mest korrekt at benytte *t*-fordelingen, hvilket statistikprogrammer også gør som det automatiske valg.

EKSEMPEL 4: SIKKERHEDSINTERVAL MED GENNEMSNIT VED BENYTTELSE AF T-FORDELING

Der er også ved estimering via t-statistik fuld overensstemmelse mellem sikkerhedsinterval og hypotesetest, når det angår gennemsnit på intervalskalerede variabler. Derfor skal der igen blot benyttes samme grundformler som ved approksimation til normalfordeling. Igen er eneste forskel, at z skiftes ud med *t*. Derfor kan et sikkerhedsinterval ved brug af *t*-værdi opstilles på følgende facon:

$$\bar{y} \mp t(se) \qquad \text{(FORMEL 7)}$$

Skal formlen gøres mere specifik, og vil man fx præsentere formlen for et 95 pct. sikkerhedsinterval, kan man *ikke* blot skrive en bestemt t-værdi, som det er tilfældet med z-approksimation. Hvor langt ud på hver side af stikprøvegennemsnittet, man skal gå, afhænger nu ikke alene af valget af sikkerhedsniveau, men også af antal frihedsgrader, dvs. stikprøvestørrelsen minus 1. Det er kutyme at benytte en notation som nedenstående, hvor det som subscript angives, at det skal være *t*-vædien gældende for en andel på 0,025 (2½ pct.) i højre hale, og hvor det eksakte tal er afhængigt af frihedsgraderne:

$$\bar{y} \mp t_{0,025}(se)$$

Fra Boks 2 fremgår, at *t*-værdien med fire decimalers nøjagtighed er lig med 1,9680, hvilket giver følgende 95 pct. sikkerhedsinterval:

$$5{,}3266 \pm 1{,}9680(0{,}1366) = 5{,}3266 \pm 0{,}2688$$

Eller skrevet som et interval med firkantede parenteser:

$$[5{,}0578;\ 5{,}5954]$$

Med 95 pct. sikkerhed kan det altså konkluderes, at populationsgennemsnittet på politisk selvplacering blandt 16-29-årige ligger mellem cirka 5,06 og 5,60, dvs. en anelse mod højre fra midten af den politiske skala, men altså i hvert fald statistisk signifikant forskelligt fra midten på et 0,05-niveau, og jo i overensstemmelse med den statistiske hypotesetest ovenfor.

I artiklen er der indtil nu alene set på sikkerhedsintervaller og hypotesetest for variabler, der er intervalskalerede, og hvor det altså giver mening at beregne gennemsnitsværdier hen over en række analyseenheder. Men mange variabler er jo enten ordinalt eller nominalt skalerede, og på den slags kategoriske variabler giver det ikke mening at beregne gennemsnit[10]. I stedet vil man typisk se på andel/proportion i en bestemt kategori eller i flere sammenlagte kategorier. Det forholder sig heldigvis sådan, at man ved hjælp af samme hovedprincipper som ved gennemsnit på intervalskalerede variabler kan estimere sikkerhedsintervaller og foretage statistisk hypotesetest for proportion. Herom handler nedenstående afsnit 5.

5. Sikkerhedsintervaller og hypotesetest for proportioner

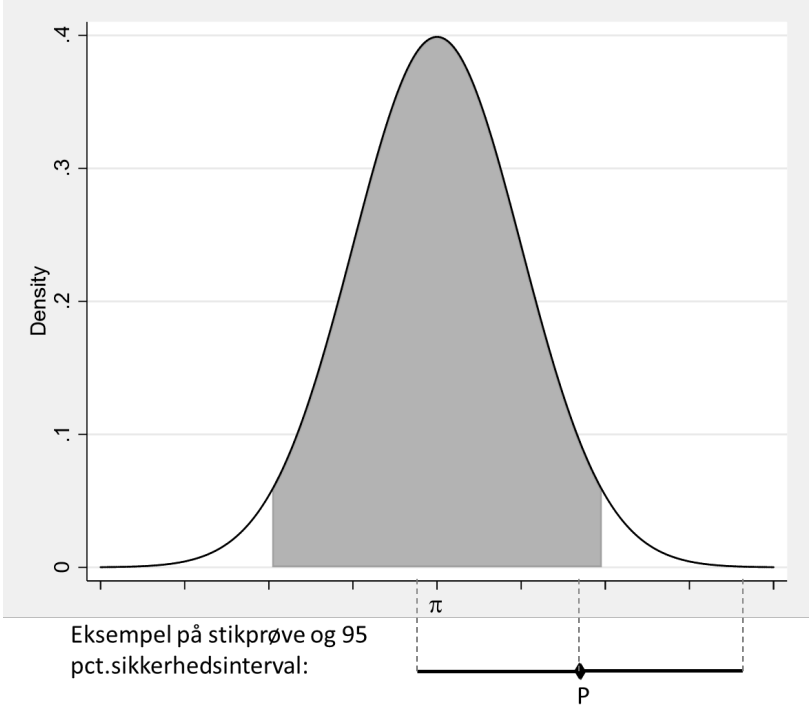
En proportion, eller med et andet ord en andel, angives med et tal mellem 0 og 1[11]. En andel på 0,5 angiver fx, at halvdelen af analyseenhederne er i pågældende kategori (eller samling af kategorier), mens den anden halvdel ikke er heri. Hvis vi fx har ti analyseenheder, og de fem er i en bestemt kategori, så er andelen i pågældende kategori 0,5, og var det kun i én af de ti analyseenheder, ville andelen være 0,1. Hvis man forestiller sig, at vi har at gøre med en såkaldt dummy-variabel, der kan antage værdierne 0 og 1, og de fem af enhederne har værdien 1, så vil ikke blot andelen, men også gennemsnittet på variablen være lig med 0,5. På tilsvarende vis vil gennemsnittet være lig med 0,1, hvis det kun var én ud af de ti analyseenheder, der havde værdien 1 på variablen. Det vil sige, at en andel er en slags gennemsnit, og det betyder faktisk, at man kan benytte samme overordnede procedure for sikkerhedsintervaller og statistisk

[10] Ved visse ordinalskalerede variabler kan man argumentere for, at de er tilnærmelsesvis intervalskalerede, hvorfor det så også giver mening at beregne gennemsnitsværdier, men ved en lang række ordinalskalerede variabler er dette ikke en farbar vej.

[11] Eller, hvis man taler om procentandele, mellem 0 og 100.

hypotesetest som ved gennemsnit på intervalskalerede variabler, sådan som det eksemplificeres ved sikkerhedsinterval-estimering i Figur 9. Eneste forskel fra Figur 4 er, at termer for gennemsnit i henholdsvis population og stikprøve er skiftet ud med termer for andel i henholdsvis population og stikprøve.

Figur 9 Skitseret samplingfordeling samt 95 pct. sikkerhedsinterval omkring stikprøveandel



Formlen for sikkerhedsinterval for andel ser fx sådan her ud, hvor "P" står for proportion/andel i stikprøve:

$$P \mp z(se), \quad \text{hvor } se = \sqrt{\frac{P(1-P)}{n}} \qquad \text{(FORMEL 8)}$$

Eneste ændring fra formlen for sikkerhedsinterval for gennemsnit ved brug af z-værdi (se Formel 5) er, at $\bar{y}$ er skiftet ud med *P*. Læg imidlertid mærke til, at formlen for den estimerede standardfejl er ændret. Det hænger sammen med, at der ved andele er sammenhæng mellem andelens størrelse og standardafvigelse, og at man derfor alene på baggrund af viden om andelens størrelse kan beregne sig direkte frem til standardafvigelsen uden at løbe samtlige observationer igennem. Standardafvigelsen, *s*, for en andel er lig med $\sqrt{(P(1-P)}$.

Proceduren for hypotesetest for andele er ligeledes overensstemmende med hoved-principperne for hypotesetest for gennemsnit, som det fremgår af formel 9 nedenfor.

$$z = \frac{P - P_0}{se_0}, \qquad \text{hvor } se_0 = \sqrt{\frac{P_0(1-P_0)}{n}} \qquad \text{(FORMEL 9)}$$

Også her beregnes afstanden mellem stikprøvens estimat og nulhypotesens værdi, udtrykt i antal standardfejl. Der er imidlertid én afgørende forskel. Læg mærke til, at der i brøkens nævner ikke blot står *se*, men se_0. Betegnelsen se_0 står for den standardfejl, der er gældende for andelen ifølge nulhypotesen. Det er dén, der er den mest korrekte i situationen, da det jo er nulhypotesen, man tester. Approksimationen til et normalfordelt z testværdi ville blive ringere hvis man i stedet for benyttede standardfejlen estimeret ud fra stikprøvens beregnede andel. Denne forskel i beregning af standardfejl mellem test og sikkerhedsinterval gør i øvrigt, at der ikke på samme måde som ved gennemsnit på intervalskalerede variabler er fuld overensstemmelse mellem sikkerhedsinterval og statistisk hypotesetest. I sjældne tilfælde kan man fx afvise en nulhypotese med alfa-værdi på 0,05, men som samtidigt ligger inden for et 95 pct. sikkerhedsinterval. Det vil dog kun være marginalt.

EKSEMPEL 5: SIKKERHEDSINTERVAL OG HYPOTESETEST FOR ANDELE

I en politisk *Voxmeter*[12] fra 16. december 2018, som ifølge meningsmålingsinstitutionen er udtrukket simpelt tilfældigt blandt danskere på 18 år og derover, får rød blok støtte fra 51,2 pct. af de 1.026 respondenter, hvis der var folketingsvalg den følgende dag. Hvis man går ud fra, at oplysningerne om repræsentativitet er korrekte, hvad kan man så på den baggrund konkludere vedrørende andelen af stemmer på partier fra rød blok i populationen blandt 18+ årige danskere?[13]

Jeg laver til en start en statistisk hypotesetest for, om man med stor grad af sikkerhed kan konkludere, at der i befolkningen på undersøgelsestidspunktet var flertal for venstrefløjspartier. Det kan jeg gøre ved at opstille en nulhypotese, der siger, at andelen i befolkningen, der vil stemme på venstrefløjspartier, er lig med 0,5, altså 50 pct. Hvis jeg kan afvise dén hypotese på baggrund af stikprøvedataene, kan jeg konkludere, at der med en vis statistisk sikkerhed er et flertal til venstre. Nul- og alternativ hypotese kan altså opstilles sådan her:

[12] https://voxmeter.dk/meningsmalinger/

[13] Selvom denne artikel først og fremmest drejer sig om de grundlæggende aspekter vedrørende statistisk inferens, skal man være opmærksom på, at der findes andre usikkerhedsaspekter end de her behandlede. Hvis mit ærinde med testen for fordeling mellem venstre og højre fx er at sige noget om fordeling af stemmer ved et potentielt folketingsvalg på undersøgelsestidspunktet, bør jeg også være opmærksom på, at der er en del respondenter, der ikke har villet svare på spørgsmålet om partivalg, og hvis disse respondenter fx langt overvejende ville stemme på partier fra den ene af de to fløje ved et rigtigt folketingsvalg, så ville min opgjorte andel ikke stemme overens hermed. Der ville være systematisk skævhed i bortfaldet. Formelt set kan jeg alene udsige mig om den population af voksne danskere, der vil besvare sådan et spørgsmål i en survey, og dette forudsætter stadigvæk også korektheden af, at stikprøven er simpelt tilfældigt udvalgt fra den samlede population.

$$H_0: \pi = 0.5$$

$$H_a: \pi \neq 0.5$$

Jeg går ud fra formel 9, hvori jeg indsætter tallene fra nulhypotese og Voxmeter-undersøgelsen:

$$z = \frac{0{,}512 - 0{,}5}{\sqrt{\frac{0{,}5(1-0{,}5)}{1026}}} = \frac{0{,}512 - 0{,}5}{\sqrt{\frac{0{,}25}{1026}}} = \frac{0{,}512 - 0{,}5}{0{,}016} = 0{,}769$$

Med dobbeltsidet test giver det en p-værdi på 0,442, altså langt fra statistisk signifikant på 0,05 niveau[14]. Nulhypotesen om et fifty fifty-fordeling mellem rød og blå blok kan altså ikke afvises på baggrund af stikprøvens data. Hvis man i stedet for vil have et mere præcist billede af, inden for hvilket interval andelen for rød blok i populationen ligger med fx 95 pct. statistisk sikkerhed, kan man estimere et 95 pct. sikkerhedsinterval ud fra formel 8:

$$0{,}512 \mp 1{,}96\sqrt{\frac{0{,}512(1-0{,}512)}{1026}}$$

$$= 0{,}512 \mp 1{,}96 \times 0{,}016$$

$$= 0{,}512 \mp 0{,}031$$

$$= [0{,}481; 0{,}543]$$

Med 95 pct. sikkerhed ligger andelen af stemmer på partier fra rød blok i populationen blandt 18+ årige danskere altså mellem cirka 48 og 54 pct. Ligesom ved den statistiske hypotesetest kan man heller ikke her konkludere, at der er flertal i befolkningen til venstrefløjspartier. Hvis det skulle have været tilfældet, ville det kræve, at sikkerhedsintervallet udelukkende var med værdier over 0,5.[15]

DER BENYTTES *IKKE* T-VÆRDIER VED TEST OG SIKKERHEDSINTERVALLER FOR ANDELE

Der skal til slut kort nævnes nogle afgørende forskelle mellem inferens for gennemsnit og andel. En andel er begrænset til at kunne antage værdier mellem 0 og 1, og hvis en andel er tæt på en af disse grænser, vil samplingfordelingen ikke kunne være

[14] Se eksempel 2 for lignende, men mere fyldig, gennemgang af eksempel på hypotesetest for gennemsnit på intervalskaleret variabel.

[15] Læg i øvrigt mærke til, at det statistisk set er ligegyldigt, om man vælger at fokusere på den ene eller anden kategori, andel stemmer på rød eller blå blok. Standardfejlen vil både i hypotesetest og sikkerhedsinterval være ens uanset valg.

tilnærmelsesvis normalfordelt, med mindre at der udtrækkes store stikprøver, og jo nærmere man kommer en af grænserne, des større stikprøve skal der udtrækkes, førend man kan benytte approksimation til standard-normalfordelt teststatistik. Endvidere vil samplingfordelingen fra udtrækning af meget små stikprøver slet ikke kunne ligne normalfordelinger, uanset hvor på skalaen mellem 0 og 1 andelen befinder sig. Udtrækkes der fx stikprøver på tre enheder i hver, vil alene andelene 0, $1/3$, $2/3$ og 1 være mulige udfald på den enkelte stikprøves andel, hvorfor samplingfordelingen ikke vil kunne være "smooth". Det giver derfor heller ikke mening at skifte til t-fordeling ved små stikprøver, sådan som man ville gøre ved inferens af gennemsnit. Af disse årsager benytter man ved små stikprøver helt andre metoder til estimering af sikkerhedsintervaller og hypotesetest, fx såkaldte *eksakt test*. Agresti (2008) anbefaler, at man ved benyttelse af ovenfor gennemgåede metode for sikkerhedsinterval for andele sætter et minimumskrav til stikprøven, så der er mindst 15 enheder i begge kategorier. Og når det angår hypotesetest, anbefaler Agresti, at det på baggrund af nulhypotesen forventede antal bør være mindst ti i hver kategori, dvs. at for en nulhypotese på 0,5 bør stikprøven være på mindst 20 enheder.

6. Afslutning

I artiklen har jeg gennemgået nogle basale aspekter vedrørende statistisk inferens fra stikprøve til population. Man kan sige, at artiklen stopper, hvor det ellers begynder at kunne blive rigtig interessant, og hvor man ser på forskellige typer af sammenhænge mellem variabler, foretager statistisk kontrol, begynder at diskutere effekter osv. Artiklen stopper også, førend det grundlæggende stof begynder at blive vanskeligt, fx i situationer hvor stikprøveudvælgelsen ikke er simpel tilfældig, eller hvor der er et muligt skævt bortfald, og hvor det ofte vil være tilrådeligt at vægte data og korrigere signifikansestimeringen ift. de simple formler, der er præsenteret igennem artiklen. Mit argument for artiklens begrænsede indhold er imidlertid, at en vis forståelse for det grundlag, som alt det vanskelige og "sjove" bygger på, er vigtigt. Det er fx vigtigt for den studerende, at han eller hun formår at formulere de statistiske konklusioner på en korrekt facon. Det er modsat meget uheldigt, når en studerende brillerer med avancerede statistiske analysemetoder, men derpå viser manglende forståelse for, hvad en p-værdi egentlig betyder. Og forståelsen betyder mere end blot en karakter til en eksamen. Der er en langt større tilfredshed i at skrive, at noget er statistisk signifikant, hvis man er nogenlunde klar over, hvad statistisk signifikans betyder, og man er samtidig i stand til at formulere ens konklusioner med langt større pondus og ligefremhed, end hvis man fx blot ser alfa-værdier som magiske grænser, man skal forsøge at kommer under. Det er jo ikke, fordi alle studerende på samfundsvidenskabelige uddannelser skal være halvvejs-statistikere eller jonglere rundt med statistiske formler på niveau med økonomistuderende, men et vist niveau af forståelse for grundlaget og logikken i statistik er en ganske sund ballast for alle samfundsvidenskabelige studerende.

Bilag 1. Notation for benyttede statistiske begreber

μ	Gennemsnit i population (my)
σ	Standardafvigelse i population (sigma)
$\bar{y}$	Gennemsnit i stikprøve (y-streg)
s	Standardafvigelse i stikprøve
$\sigma_{\bar{y}}$	Standardfejl for gennemsnit (sigma y-streg)
se	Estimeret standardfejl
π	Andel/proportion i population
P	Andel/proportion i stikprøve
z	Teststørrelse ved andele samt ved gennemsnit med kendt standardafvigelse i population
t	Teststørrelse ved gennemsnit

Formler for $\bar{y}$ og s:

Netop størrelserne "gennemsnit" og "standardafvigelse" er noget af det eneste, der i artiklen forudsættes kendskab til på forhånd, men for god ordens skyld vises her formler for disse to, sådan som de beregnes ud fra en stikprøve.

Gennemsnit på en variabel *y* i en stikprøve: $\bar{y} = \frac{\sum y_i}{n}$, hvor *n* er antal observationer i stikprøven.

Gennemsnittet er altså lig med summen af samtlige værdier på variablen divideret med stikprøvestørrelsen.

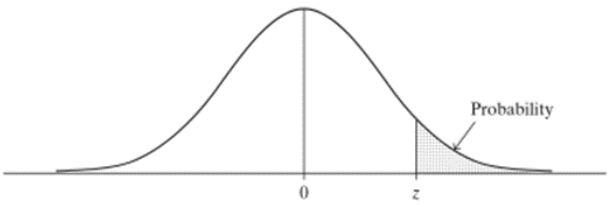
Standardafvigelse på en variabel y ud fra en stikprøve: $s = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}}$

Standardafvigelsen er altså lig med kvadratroden af en brøk, hvor tælleren er summen af kvadrerede afvigelser mellem den enkelte observation og stikprøvegennemsnittet, og hvor nævneren er stikrøvestørrelsen minus 1. Ved at fratrække 1 i nævneren fås et mere validt estimat af standardafvigelsen i populationen.

Variansen er den størrelse, der står inde under kvadratrodstegnet, dvs. s^2

Bilag 2. Standardnormalfordelingstabel med angivelse af sandsynligheder i højre hale[16]

TABLE A: Normal curve tail probabilities. Standard normal probability in right-hand tail (for negative values of z, probabilities are found by symmetry).



	Second Decimal Place of z									
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0352	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.00135									
3.5	.000233									
4.0	.0000317									
4.5	.00000340									
5.0	.000000287									

Source: R. E. Walpole, *Introduction to Statistics* (New York: Macmillan, 1968).

[16] Tabellen er taget fra Agresti (2018) p561, men som det fremgår under tabellen, er der en oprindelig kilde, R. E. Walpole (1968).