

Sandsynlighedsregning og statistik med binomialfordelingen

Katja Kofod Svan og Olav Lyndrup

Januar 2019

Binomialtest

Binomialfordelinger anvendes til at udtage stikprøver *med* tilbagelægning – eller stikprøver, der udtages af meget store populationer, hvor det spiller en ubetydelig lille rolle, om der tilbagelægges eller ej. Skal der derimod udtages en stikprøve *uden* tilbagelægning, så skal vi have fat på en anden sandsynlighedsfordeling. Denne kaldes den hypergeometriske fordeling. Vi vil ikke gå dybere ind i eksperimenter med og uden tilbagelægning, men for interesserede behandles den hypergeometriske fordeling i appendix sidst i dette materiale.

Indtil nu har vi udtaget stikprøver fra en kendt mængde, og vi har ud fra forskellige sandsynlighedsmodeller udtalt os om sammensætningen af stikprøven. I dette afsnit vil vi på baggrund af konkrete stikprøver i stedet for udtale os om den mængde, en konkret stikprøve er udtaget fra.

Når vi gør dette, så formulerer vi på forhånd en formodning, som vi vil undersøge. En sådan formodning kaldes en *nulhypotese*, og den formuleres som oftest som et udsagn om, at den variation, vi observerer, er udtryk for tilfældig variation i den givne situation. Det modsatte udsagn til en nulhypotese kaldes *den alternative hypotese*. Den alternative hypotese udtrykker således, at den observerede variation er udtryk for systematiske afvigelser fra det, man normalt ville observere i den givne situation.

Når vi udfører *et hypotesetest*, antager vi, at *nulhypotesen er sand*, og vi udfører hypotesetestet med udgangspunkt i denne antagelse. Hvis man i hypotesetestet når frem til noget meget usandsynligt, dvs. hvis vi får en meget lille sandsynlighed for, at det vi undersøger indtræffer, så konkluderer vi, at nulhypotesen må forkastes. Vi fastsætter selv grænsen for, hvor lille en sandsynlighed vi vil acceptere, før vi siger, at resultatet er så usædvanligt, at vores nulhypotese ikke længere er troværdig. Når vi forkaster en nulhypotese, så svarer

det til, at vi i stedet for at tro på nulhypotesen foretrækker at tro på den alternative hypotese om systematisk variation. En konklusion på et statistisk hypotesetest rummer således altid et subjektivt element.

Det niveau, vi fastsætter som grænsen mellem de resultater, vi accepterer under antagelse af, at nulhypotesen er sand, og de resultater, vi finder så usædvanlige, at nulhypotesen ikke længere kan opretholdes, kaldes for *signifikansniveauet*. Signifikansniveauet kan ikke begrundes matematisk. I praksis bruges ofte 5% som signifikansniveau, men man kan møde problemer, fx i medicinske forsøg eller erhvervskontrakter, hvor dette fastsættes til 1% eller 10%. Hvis vi fastsætter et højt signifikansniveau, fx på 10%, så bliver det lettere at forkaste nulhypotesen, og hvis vi omvendt fastsætter et lavt signifikansniveau, betyder det, at vi accepterer lidt flere usædvanlige hændelser, før vi forkaster nulhypotesen. Jo mindre signifikansniveauet er, jo sværere er det altså at forkaste nulhypotesen om, at de observerede afvigelser alene beror på uundgåelige tilfældige variationer.

Eksempel 14 Møntkast

Jesper vil undersøge en mønt, der skal bruges til lodtrækning. Han kaster derfor mønten fem gange og opdager, at den lander på ”krone” alle fem gange. Dette virker mistænksomt, men kan han nu slutte, at mønten er uærlig?

For en ærlig mønt vil vi forvente gennemsnitlig lige mange ”krone” og ”plad”, dvs. sandsynligheden for at få ”krone” er.

Jesper antager, at de 5 kast med mønten kan modelleres med en binomialfordelt stokastisk variabel X , som tæller antallet af krone. Antalsparameteren n er så 5. Vi bemærker, at den observerede værdi af X her også er 5. Den observerede værdi i et hypotesetest kaldes også *teststørrelsen*.

Han opstiller nulhypotesen H_0 og den alternative hypotese H_1 , idet sandsynligheden for succes, dvs. sandsynligheden for at få ”krone”, betegnes p :

$$H_0: \text{Mønten er ærlig, dvs. } p = \frac{1}{2}.$$

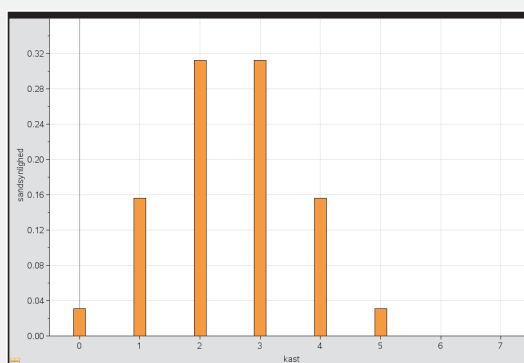
$$H_1: \text{Mønten er uærlig, dvs. } p \neq \frac{1}{2}.$$

Signifikansniveauet fastsætter vi til 5%.

I dette hypotesetest forkastes nulhypotesen både, hvis der kastes for mange og for få ”krone”. Denne type hypotesetest kaldes *to-sidet*, fordi vi smider væk fra begge sider i diagrammet – begge haler.

De 5% skal derfor fordeles med 2,5% i hver side af fordelingen. De 95% midterste af udfaldene kaldes *acceptområdet*, mens de to haler, som udgør sidste 5% af udfaldene kaldes *det kritiske område*.

Binomialmodellen $b(5, \frac{1}{2})$ giver os følgende sandsynlighedsfordeling repræsenteret ved et søjlediagram:



Af søjlediagrammet fremgår det, at sandsynligheden for at få 5 ”krone” i 5 kast er større end 2,5%, og nulhypotesen kan derfor ikke forkastes i dette tilfælde.

Sandsynligheden for at få 5 ”krone” i 5 kast kan også beregnes som:

$$P(X = 5) = \binom{5}{5} \cdot 0,5^5 \cdot 0,5^{5-5} = \frac{5!}{5!0!} \cdot 0,5^5 \cdot 0,5^0 = 0,5^5 = 0,03125.$$

Det betyder, at samtlige udfald i eksperimentet ligger inden for acceptområdet. Vi kan skrive acceptområdet som mængden, der indeholder tallene $\{0,1,2,3,4,5\}$. Dvs. uanset, hvilket udfald vi får af de 5 kast, så vil det ikke være usædvanligt.

Definition 8 Tosidet binomialtest

En binomialfordelt stokastisk variabel X har antalsparameteren n .

Der er formuleret en nulhypotese, som udtrykker en formodning om, at sandsynlighedsparameteren p netop har en bestemt værdi:

$$H_0: p = p_0$$

Acceptområdet betegnes $\{k_v, \dots, k_h\}$, hvor k_v betegner det mindste acceptable udfald (dvs. det første acceptable udfald set fra venstre) og k_h betegner det største acceptable udfald (dvs. det første acceptable udfald set fra højre). Værdierne k_v og k_h beregnes ud fra et signifikansniveau på 5%, så der er 2,5% i hver side:

Værdien k_v bestemmes som den mindste værdi af X , så $P(X \leq k_v) > 0,025$.

Værdien k_h bestemmes som den største værdi af X , så $P(X \geq k_h) > 0,025$.

Det undersøges om den *observerede værdi* af X , også kaldet *teststørrelsen*, ligger i acceptområdet eller ej.

H_0 accepteres, hvis teststørrelsen ligger i acceptområdet.

H_0 forkastes, hvis teststørrelsen ligger udenfor acceptområdet.

Eksempel 15 Møntkast (fortsat)

Hvis vi i eksemplet med møntkast i stedet vil undersøge, om der i de 5 kast kommer for mange af den ene slags udfald, så formuleres nulhypotesen lidt anderledes, og konklusionen kan blive en anden. Lad os antage, at vi har en mistanke om, at mønten giver for mange ”krone”.

Hertil opstiller vi nulhypotesen og den alternative hypotese som følger:

H_0 : Mønten giver ”krone” i højst halvdelen af tilfældene, dvs. $p \leq \frac{1}{2}$.

H_1 : Mønten giver ”krone” i flere end halvdelen af tilfældene, dvs. $p > \frac{1}{2}$.

Signifikansniveauet er stadig 5% og sandsynlighedsfordelingen er den samme som ovenfor, dvs. $b(n, p) = b(5, \frac{1}{2})$. I dette tilfælde er testet dog *ensidet*, da vi ikke forkaster H_0 , hvis der er for få ”krone”, dvs. for mange ”plat”. Det kritiske område ligger derfor samlet i ”højre” side af fordelingen.

Vi har stadig, at $P(X = 5) = 0,03125$, og da en udvidelse med 4 ”krone” giver en sandsynlighed langt over signifikansniveauet:

$$P(X \geq 4) = P(X = 4) + P(X = 5) = 0,1563 + 0,03125 = 0,1875 > 5\%$$

kan vi konkludere, at acceptområdet består af udfaldene 0, 1, 2, 3 og 4 "krone" i 5 kast. Det kritiske område består således kun af det udfald, der giver 5 "krone" i 5 kast.

I undersøgelsen landede mønten, som blev kastet fem gange, hver gang på "krone". Den observerede værdi ligger altså ikke i acceptområdet, og nulhypotesen forkastes derfor. Konklusionen er, at vi foretrækker den alternative hypotese, nemlig at mønten giver "krone" i flere end halvdelen af tilfældene. Da det kritiske område ligger til "højre", når de mulige udfald for X opskrives $\{0,1,2,3,4,5\}$, så kaldes binomialtestet for *højresidet*.

Definition 9 Højresidet binomialtest

En binomialfordelt stokastisk variabel X er givet med antalsparameteren n .

Der er formuleret en nulhypotese, som udtrykker en formodning om, at sandsynlighedsparameteren p er *mindre end* en bestemt værdi:

$$H_0: p \leq p_0.$$

Acceptområdet betegnes $\{0, \dots, k_h\}$, hvor k_h betegner det største acceptable udfald (dvs. det første acceptable udfald set fra højre).

Værdien k_h beregnes ud fra et signifikansniveau på 5% med de 5% liggende til højre i fordelingen:

$$\text{Værdien } k_h \text{ bestemmes som den største værdi af } X, \text{ så } P(X \geq k_h) > 0,05.$$

Det afgøres om den *observerede værdi* af X , dvs. *teststørrelsen*, ligger i acceptområdet eller ej.

H_0 accepteres, hvis teststørrelsen ligger i acceptområdet.

H_0 forkastes, hvis teststørrelsen ligger udenfor acceptområdet.

Opgave 22 Formuler en definition for et venstresidet hypotesetest i binomialfordelingen.

Opgave 23 Søren vil undersøge en mønt, der skal bruges til lodtrækning. Han kaster mønten et antal gange og får mistanke om, at den lander på krone for sjældent.

a) Opstil en nulhypotese for et venstresidet binomialtest.

Han kaster nu mønten otte gange og observerer, at den lander på "krone" to gange.

b) Afgør med et 5% signifikansniveau, om nulhypotesen kan forkastes eller ej.

Opgave 24 I et casino spilles der med en otte-sidet terning med et af tallene 1 til 8 på hver af siderne. Dvs. der er 8 forskellige udfald ved et kast med denne terning, og hvert udfald indtræffer med sandsynligheden $\frac{1}{8}$. En gæst kaster terningen 10 gange og observerer, at der ikke blev slået en eneste otter.

a) Opstil en nulhypotese, og benyt et tosidet binomialtest med et 5% signifikansniveau til undersøge, om terningen er ærlig.

- b) Opstil en nulhypotese, og benyt et ensidet binomialtest med et 5% signifikansniveau til undersøge, om sandsynlighedsparameteren er større end $\frac{1}{8}$.

Opgave 25 Undersøg, hvordan dit værktøjsprogram udfører binomialtest

I definition 8 og 9 er nulhypotesen accepteret eller forkastet ved først at bestemme acceptmængden og dermed kritisk mængde, og derefter en afgøre om teststørrelsen, dvs. den observerede værdi, ligger i acceptmængden eller ej.

I stedet for at se på teststørrelsens værdi, kan vi afgøre hypotesetest ud fra en *p-værdi*. Bemærk, at *p*-værdien ikke må forveksles med basissandsynligheden *p*. I et binomialtest angiver *p*-værdien sandsynligheden for at få et udfald, der er mindst lige så skævt som det aktuelle, når det antages, at nulhypotesen er sand. I eksemplet ovenfor med det to-sidede test med mønten fik vi en *p*-værdi på $2 \cdot 0,03125 = 6,25\%$. *p*-værdien beregnes således ud fra teststørrelsen, som en kumuleret sandsynlighed, hvor vi udnytter vores viden om, hvorvidt testet er en- eller tosidet.

For at afgøre om nulhypotesen skal forkastes eller ej sammenholdes *p*-værdien med signifikansniveauet. Følgende eksempel viser, hvor man afgør et hypotesetest ud fra *p*-værdien.

Eksempel 16 Terningekast

En firesidet terning med et af tallene fra 1 til 4 på hver side kastes 20 gange, og det observeres, at antallet af toere er 3. Dvs. der er 4 forskellige udfald ved et kast med denne terning, og hvert udfald indtræffer med sandsynligheden $\frac{1}{4}$.

Vi vil undersøge nulhypotesen, herunder den alternative hypotese:

$$H_0: \text{Terningen giver toere en fjerdedel af gangene, dvs. } p = \frac{1}{4}.$$

$$H_1: \text{Terningen giver ikke toere en fjerdedel af gangene, dvs. } p \neq \frac{1}{4}.$$

Det antages, at antallet af toere kan modelleres med en binomialfordelt stokastisk variabel *X*, der tæller antallet af toere. Antalsparameteren *n* er 20.

Vi vil undersøge nulhypotesen med et to-sidet hypotesetest, hvor vi fastsætter signifikansniveauet til 5%, dvs. 2,5% i hver side.

Den observerede værdi af *X*, dvs. teststørrelsen, er 3. Ud fra teststørrelsen beregner vi *p*-værdien (sandsynligheden for at få noget, der er mindst lige så skævt som det observerede) ved først at bestemme den kumulerede sandsynlighed: $P(X \leq 3) = 0,225$.

Da testet er to-sidet bestemmes *p*-værdien som: $2 \cdot 0,225 = 45\%$.

Da *p*-værdien er (meget) større end signifikansniveauet på 5% kan vi ikke forkaste vores nulhypotese, dvs. det tyder på, at den firesidede terning giver toere i en fjerdedel af gangene.

I stedet for at benytte værktøjsprogrammernes indbyggede sandsynlighedsfordelinger, kan man også benytte værktøjsprogrammerne til at *simulere* nulhypoteserne.

Eksempel 17 Blindsmagning

Et slikfirma ønsker at undersøge, om der er forskel på smagen af gule og røde vingummier. De foretager derfor en såkaldt triangeltest, hvor en række forsøgspersoner får forelagt tre vingummier, hvoraf to af dem er gule og den sidste er rød, eller omvendt. I en smagning smager forsøgspersonerne de tre vingummier uden at se farven på dem og vælger til sidst, hvilken af de tre der er forskellig fra de andre.

Vi lader den stokastiske variabel X tælle antallet af forsøgspersoner, der vælger ”den rigtige”.

Vi har på forhånd en forventning om, at man ikke kan smage forskel, dvs. en tilfældig tredjedel af forsøgspersonerne vil vælge ”den rigtige” eller en af de ”to forkerte”. Vi lader derfor hypotesetesten være ensidet, nemlig højresidet, svarende til, at højst en tredjedel af forsøgspersonerne vil vælge ”den rigtige”.

Nullhypotesen er derfor, at højst en tredjedel af forsøgspersonerne kan smage forskel i hver af smagningerne, og den alternative hypotese er dermed, at mere end en tredjedel af forsøgspersonerne kan smage forskel i hver af smagningerne. Lader vi p betegne sandsynligheden for, at en forsøgsperson kan smage forskel, er vores nullhypotese og alternative hypotese:

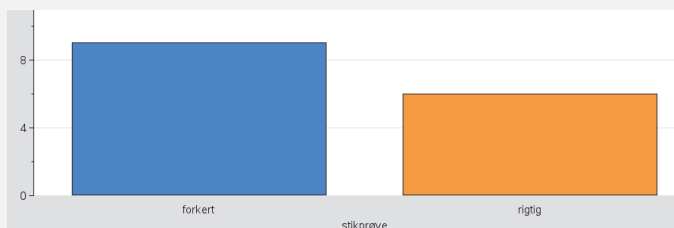
$$H_0: p \leq \frac{1}{3}. \quad H_1: p > \frac{1}{3}.$$

Hvis vi lader 15 forsøgspersoner foretage blindsmagningen, som beskrevet, så vil vi forvente, at højst 5 af dem kan smage forskel på vingummierne.

I en konkret undersøgelse vælger 8 af forsøgspersonerne ”den rigtige” vingummi i blindsmagningen, mens 7 ikke kan. Teststørrelsen, dvs. vores observerede værdi, er altså 8.

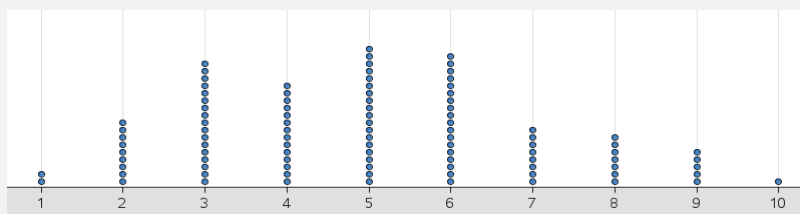
For at undersøge, om det er et udtryk for tilfældig variation, simulerer vi nu situationen i stedet for at foretage teoretiske beregninger i binomialfordelingen.

Vi lader værktøjsprogrammet vælge en af de tre vingummier på tilfældig måde 15 gange svarende til de 15 forsøgspersoners valg i en blindsmagning. Vi får så fx en *stikprøve*, hvor 9 testpersoner valgte en af de ”to forkerte”, og 6 valgte ”den rigtige”, som vist i søjlediagrammet herunder.

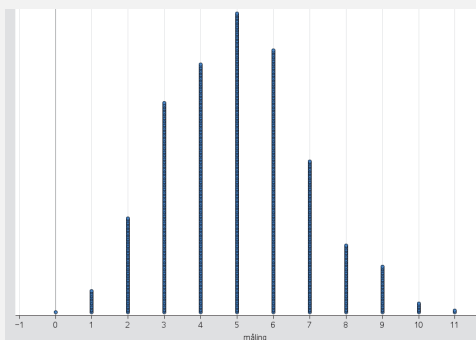


Vi lader nu værktøjsprogrammet udtage en sådan stikprøve med 15 forsøgspersoner rigtig mange gange, og vi noterer hver gang, hvor mange af forsøgspersonerne, der vælger ”den rigtige” vingummi.

Efter simulering af 100 stikprøver kunne fordelingen af det antal forsøgspersoner, der valgte ”den rigtige” vingummi, se sådan ud:



Efter simulering af 1000 stikprøver kunne fordelingen se sådan ud:



Det bemærkes, at fordelingen som forventet efterhånden kommer til at ligne binomialfordelingen $b(15, \frac{1}{3})$ svarende til sandsynlighedsfordelingen for en stokastiske variabel X med antalsparameter $n = 15$ og sandsynlighedsparameter $p = \frac{1}{3}$.

I 90 ud af 1000 simulerede stikprøver fandt vi, at 8 eller flere af testpersonerne valgte ”den rigtige” vingummi. Det giver derfor en sandsynlighed på 9% for at få en stikprøve, der er mindst lige så skæv som den observerede, når det forudsættes, at nulhypotesen er sand. De 9% kaldes også for den *eksperimentelle p-værdi*.

Med et signifikansniveau på 5% kan vi derfor ikke forkaste nulhypotesen om, at man højst kan smage forskel i en tredjedel af smagningerne. Eller sagt på en anden måde: Der er ikke belæg for at påstå, at man kan smage forskel på en gul og en rød vingummi.

Opgave 26

Vi ser nu på eksemplet ovenfor om blindsmagning af vingummi med en teoretisk fremfor en eksperimentel vinkel, dvs. vi ser på samme nulhypotese og binomialfordelingen $b(15, \frac{1}{3})$ med en teststørrelse på $X = 8$.

- Opstil en tabel over sandsynlighederne.
- Tegn et søjlediagram for sandsynlighedsfunktionen.
- Beregn $P(X = 8)$.
- Beregn p -værdien.
- Bestem den kritiske mængde.
- Bestem det mindste antal af forsøgspersoner i stikprøven, der skal kunne smage forskel, for at man kan forkaste nulhypotesen.

Opgave 27

Ved et valg fik et bestemt parti 30% af stemmerne. Nogen tid efter foretages en meningsmåling for at undersøge, om tilslutningen til partiet har *ændret* sig. Vi er altså interesserede i at finde ud af, om partiet er gået frem *eller* tilbage. I en stikprøve på 200 personer, der er repræsentativt udvalgt, skal de udvalgte svare på, om de ville stemme på partiet, hvis der var valg i dag.

Vi lader den stokastiske variabel X tælle antallet af de adspurgte, der vil stemme på partiet i dag.

Bemærk, at stikprøven i praksis udtages uden tilbagelægning, da man i en meningsmåling ikke vil bede den samme person svare flere gange, og populationen er i denne sammenhæng i øvrigt også meget stor i forhold til stikprøven, så der er i praksis ingen forskel på, om stikprøven foretages med eller uden tilbagelægning. I stikprøven var der 75 personer, der ville stemme på partiet, hvis der var valg i dag.

- Opstil en nulhypotese, når det antages at testet er dobbeltsidet.
- Undersøg, om man kan forkaste nulhypotesen med et 5% signifikansniveau.

Opgave 28

Et parti fik ved et valg en vælgertilslutning på 21%. En efterfølgende Gallup-undersøgelse baseret på 1008 personer viste, at partiet havde en vælgertilslutning på 18%.

Vi lader den stokastiske variabel X tælle antallet af de adspurgte, der stemmer på partiet.

- Formuler en nulhypotese, hvor udgangspunktet er, at partiets vælgertilslutning er uændret.
- Afgør om testet af nulhypotesen skal være tosidet eller ensidet.
- Opbyg en simulering af nulhypotesen i et matematisk værktøjsprogram.
- Bestem teststørrelsen, dvs. den observerede værdi.
- Viser simuleringen den samme vælgertilslutning, som Gallups undersøgelse vist? Eller er den større?
- Udfør nu 1000 simuleringer af nulhypotesen.
- Bestem antallet af simuleringer, der viser den samme eller en større værdi for vælgertilslutningen.
- Bestem den eksperimentelle p -værdi, og sammenlign med signifikansniveauet.

Konfidensintervaller

Der offentliggøres jævnligt analyser, der viser, hvilke politiske partier der går frem eller tilbage.

Chokmåling: Vælgerne flygter fra Løkke

Venstre går frem i ny Gallup

Blå blok øger sit forspring i helt ny Gallup-måling. Det er Venstre, der går betydeligt frem og sørger for, at de borgerlige kan lægge yderligere afstand til rød blok.

Ny måling: Rød blok fisker pludselig vælgere hos de blå

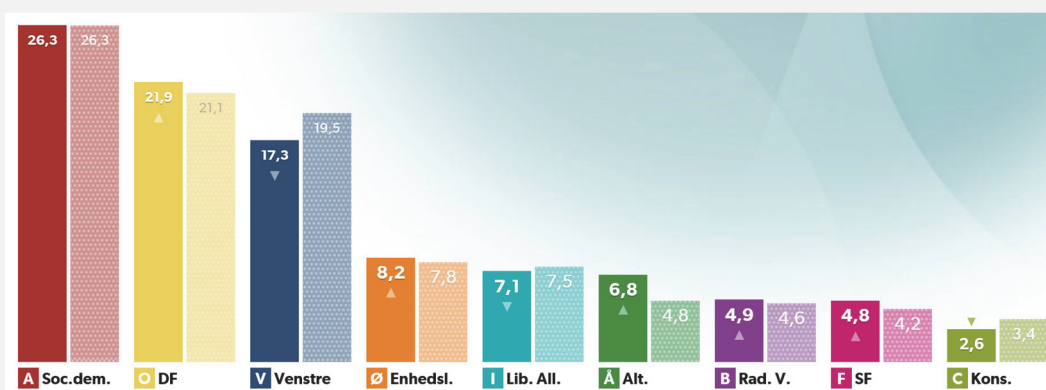
De røde partier henter mandater i den borgerlige lejr, viser ny Epinion-måling. Alternativet får 1,8 procent af stemmerne.

Disse analyser laves oftest på baggrund af telefoninterviews blandt et repræsentativt udsnit af den samlede vælgergruppe. Hvis denne repræsentative stikprøve blot er på omkring 1000 vælgere, så kan man faktisk komme ret tæt på en præcis forudsigelse, selvom der er over 4 mio. vælgere.

Det er dog klart, at der er en vis usikkerhed, når man ved at spørge så få vil forudsige, hvordan resten af billedet ser ud. Vi skal i dette afsnit se nærmere på, hvordan disse statistiske usikkerheder skal forstås, og hvordan man beregner dem.

Eksempel 18 Meningsmålinger

På figuren nedenfor ses tallene fra en meningsmåling foretaget af Epinion d. 24.8.2016 (venstre kolonne). Ligeledes ses tallene fra folketingsvalget i 2015 (højre kolonne). For eksempel kan vi se, at Venstre fik 19,5% ved valget, og at der i stikprøven kun er 17,3%, der vil stemme på dem. Er det nu et udtryk for, at Venstre er gået tilbage, eller kan den tilsyneladende nedgang forklares med den usikkerhed, der altid vil ligge i, at det ene tal stammer fra et valg, hvor "alle" deltager, mens det andet stammer fra en stikprøve, der kan have en over- eller underrepræsentation af de forskellige vælgergrupper.



Kilde: www.dr.dk

Mange analyseinstitutter noterer denne usikkerhed, når resultaterne offentliggøres. I undersøgelsen ovenfor er usikkerheden $\pm 2,5$ procentpoint og med dette som udgangspunkt, kan man i en vis forstand konkludere, at Venstre faktisk ikke er gået tilbage siden valget, fordi $17,3 + 2,5 = 19,8 > 19,5$. Ser vi på Alternativet, så står de til en fremgang på 2 procentpoint – hvad kan vi konkludere her? Og er usikkerheden den samme, uanset om vi taler om store eller små partier?

Ved valget spurgte man jo samtlige vælgere, så de procentandele, der står i søjlerne til højre, er selvfølgelig de faktiske tal. Vi ved altså, at der med sikkerhed var 19,5%, der stemte på Venstre ved folketingsvalget.

Når der nu udspørges 1578 vælgere, er det ikke rimeligt at påstå, at der er præcis 17,3%, der nu stemmer på Venstre. Og vi kan derfor heller ikke konkludere, at Venstre er gået tilbage med 2,2 procentpoint siden folketingsvalget.

Var der valg på det tidspunkt, hvor stikprøven blev indsamlet, ville dette have resulteret i en bestemt vælgertilslutning, som vi kalder *den sande værdi*. Vi kender således ikke den sande værdi, men vi ønsker at kunne udtale os om denne med en vis sikkerhed.

Forestiller vi os, at der var indsamlet tusindvis af stikprøver i samme uge, så ville det resultere i tusindvis af forskellige værdier (her procenttal), der ville ligge normalfordelt omkring den sande værdi. 95% af disse ville ligge inden for normalområdet, jævnfør afsnittet om normalfordelingen (side 13).

Omkring hvert eneste af disse stikprøve-værdier kunne vi lægge et tilsvarende interval, og for 95% af disse stikprøver ville det tilsvarende normalområde indeholde den sande værdi. Et sådant interval omkring en stikprøveværdi kaldes et *konfidensinterval*.

Definition 10 Konfidensinterval

Et 95% konfidensinterval for en estimeret parameter i en stikprøve er et interval, der opfylder, at den sande værdi for parameteren med 95% konfidens vil ligge i intervallet.

Vi siger, at konfidensintervallet indeholder den sande værdi for parameteren med 95% konfidens.

Definition 10 skal forstås sådan, at vi udtager en stikprøve, hvortil der beregnes et konfidensinterval for parameteren. Hvis vi udtager en stikprøve, uendeligt mange gange, og hver gang beregner et konfidensinterval for parameteren, vil 95% af disse konfidensintervaller indeholde den sande værdi af parameteren. Hvis vi eksempelvis i praksis gentager dette 100 gange, dvs. udtager 100 stikprøver og beregner et konfidensinterval for parameteren i hver af stikprøverne, så vil der kun være 5 af de beregnede konfidensintervaller, der *ikke* indeholder den sande værdi. Før stikprøven indsamles er der altså 95% sandsynlighed for, at man får beregnet et konfidensinterval, som faktisk indeholder parameterens sande værdi.

For at beskrive konfidensintervallet teoretisk lader vi den stokastiske variabel X tælle antallet af adspurgte i stikprøven, der vil stemme på et bestemt parti, og vi antager, at X er binomialfordelt med antalsparameter n (der svarer til stikprøvens størrelse), og sandsynlighedsparameter p (der svarer til partiets vælgertilslutning blandt hele populationen). Stikprøven udtages i praksis uden tilbagelægning, da man ikke vil ringe til den samme person to gange. Da populationen er meget stor i forhold til stikprøven, gør det i øvrigt heller ingen forskel, om stikprøven foretages med eller uden tilbagelægning.

Vi kender stikprøvens størrelse, n , mens p , som er sandsynligheden for, at en person vil stemme på et bestemt parti, er ukendt. Vi estimerer derfor sandsynligheden \hat{p} for, at en person vil stemme på et bestemt

parti ud fra stikprøven, sådan at $\hat{p} = \frac{X}{n}$, hvor X er det antal af personer i stikprøven, der vil stemme på et

bestemt parti. Hvis stikprøven er repræsentativt udvalgt, vil dette være et godt estimat. De variationer i vælgertilslutning, der vil være i forskellige stikprøver, vil hænge sammen med spredningen i binomialfordelingen. Spredningen på antallet af vælgere, der stemmer på et bestemt parti,

$\sigma = \sqrt{n \cdot \hat{p} \cdot (1 - \hat{p})}$, afhænger af stikprøvens størrelse og størrelsen af partiets vælgertilslutning, og dermed ikke af populationens størrelse. Man kan derfor få samme nøjagtighed i meningsmålinger i meget større lande uden af øge størrelsen på stikprøven.

Vi ved fra normalfordelingen (jf. opgave 21), at 95% af observationerne ligger inden for intervallet $\mu - 1,96 \cdot \sigma \leq x \leq \mu + 1,96 \cdot \sigma$. Dette interval benyttes ofte i samfundsfag, men vi vælger her at bestemme konfidensintervallet med udgangspunkt i binomialfordelingen, hvor de normale værdier for den stokastiske variabel X svarer til ca. 95% af observationerne. De normale værdier er som tidligere nævnt defineret til at ligge i intervallet $\mu - 2 \cdot \sigma \leq x \leq \mu + 2 \cdot \sigma$.

Dette betyder altså, at for ca. 95% af alle stikprøver, vil intervallet

$$n \cdot \hat{p} - 2 \cdot \sqrt{n \cdot \hat{p} \cdot (1 - \hat{p})} \leq x \leq n \cdot \hat{p} + 2 \cdot \sqrt{n \cdot \hat{p} \cdot (1 - \hat{p})}$$

indeholde den sande værdi for *antallet af personer* i stikprøven, der vil stemme på et bestemt parti.

Kigger vi i stedet på procentandele, skal vi dividere med stikprøvens størrelse. Den sande *procentandel* af vælgere, der stemmer på et bestemt parti, vil derfor for ca. 95% af alle stikprøver ligge i intervallet:

$$\hat{p} - \frac{2 \cdot \sqrt{n \cdot \hat{p} \cdot (1 - \hat{p})}}{n} \leq \frac{x}{n} \leq \hat{p} + \frac{2 \cdot \sqrt{n \cdot \hat{p} \cdot (1 - \hat{p})}}{n}$$

Vi omskriver udtrykket ved at omskrive $n = \sqrt{n^2}$ og udnytte, at $\frac{\sqrt{a}}{\sqrt{b}} = \sqrt{\frac{a}{b}}$:

$$\hat{p} - \frac{2 \cdot \sqrt{n \cdot \hat{p} \cdot (1 - \hat{p})}}{\sqrt{n^2}} \leq \frac{x}{n} \leq \hat{p} + \frac{2 \cdot \sqrt{n \cdot \hat{p} \cdot (1 - \hat{p})}}{\sqrt{n^2}}$$

$$\hat{p} - 2 \cdot \sqrt{\frac{n \cdot \hat{p} \cdot (1 - \hat{p})}{n^2}} \leq \frac{x}{n} \leq \hat{p} + 2 \cdot \sqrt{\frac{n \cdot \hat{p} \cdot (1 - \hat{p})}{n^2}}$$

$$\hat{p} - 2 \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \leq \frac{x}{n} \leq \hat{p} + 2 \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}$$

Konfidensintervallet for andelen i sådanne meningsmålinger kan altså beregnes ved:

$$\left[\hat{p} - 2 \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}; \hat{p} + 2 \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \right].$$

Dette giver os følgende sætning:

Sætning 6 Statistisk usikkerhed i stikprøver

Når der udtages en stikprøve, så bestemmes stikprøveresultatets 95% konfidensinterval ved følgende formel.

$$\left[\hat{p} - 2 \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}; \hat{p} + 2 \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \right],$$

hvor \hat{p} er den estimerede sandsynlighedsparameter, og n er antalsparameteren.

Størrelsen $2 \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}$ kaldes den *statistiske usikkerhed* eller blot *usikkerhed* i stikprøver.

Eksempel 19 Meningsmålinger (fortsat)

I vores eksempel 18 fra før, hvor $\hat{p} = 0,173$ og $n = 1578$, er 95% konfidensintervallet på partiet Venstres vælgertilslutning altså:

$$\left[0,173 - \frac{2 \cdot \sqrt{0,173 \cdot (1 - 0,173)}}{\sqrt{1578}}; 0,173 + \frac{2 \cdot \sqrt{0,173 \cdot (1 - 0,173)}}{\sqrt{1578}} \right] =$$

$$[0,173 - 0,019; 0,173 + 0,019] = [0,154; 0,192].$$

Venstres vælgertilslutning ligger altså med 95% konfidens i intervallet fra 15,4% til 19,2%, dvs. der er 95% sandsynlighed for, at dette interval rummer den sande værdi for Venstres vælgertilslutning på det pågældende tidspunkt.

Og omvendt må vi altså konkludere, at hvis vi skal tro på, at Venstres vælgertilslutning har ændret sig, så kræver det, at den vælgertilslutning, vi så ved valget, ligger uden for konfidensintervallet, dvs. den skal være mindre end 15,4% eller større end 19,2%.

Ved folketingsvalget fik Venstre 19,5% af stemmerne, altså en værdi, der netop ligger uden for konfidensintervallet, så den pågældende analyse fra Epinion fortæller os altså, at Venstre er gået tilbage.

I meningsmålinger angives ofte en usikkerhed, som i tilfældet for Venstres vælgerandel altså er $\pm 1,9$ procentpoint jævnfør beregningen i eksemplet ovenfor.

Opgave 29 Undersøg, hvordan dit værktøjsprogram beregner konfidensintervaller.

Opgave 30 Bestem usikkerheden på partiet Alternativets vælgertilslutning ud fra eksempel 18, og afgør om de er gået frem siden valget.

I følgende opgave simulerer vi et eksperiment, hvor der udtages 100 stikprøver. For hver stikprøve udregner vi konfidensintervallet, og tæller hvor mange af disse, der indeholder den sande andel.

- Opgave 31**
- Opret en binomialfordelt stokastisk variabel X i et regneark ved hjælp af en tilfældighedsgenerator. Vi sætter $n = 1$ og vælger $p = 0,21$. En sådan stokastisk variabel kaldes Bernoulli-fordelt, fordi antalsparameteren er 1.
 - Opret en stikprøve på 200 værdier af den stokastiske variabel X i regnearket.
 - Bestem middelværdien og spredningen for de 200 værdier.
 - Bestem konfidensintervallet for den sande værdi af p ud fra middelværdi og spredning.
 - Afgør om værdien 0,21 ligger i konfidensintervallet.
 - Tegn konfidensintervallet som et linjestykke i dit værktøjsprogram.
 - Gennemfør trinene a) - f) 100 gange.
 - Bestem andelen af konfidensintervaller, der indeholder værdien 0,21.
 - Hvad kan du konkludere?

Bemærk, at hvis vi ønsker at gennemføre en undersøgelse, hvor usikkerheden skal begrænses til en bestemt værdi, så kan vi ud fra sætning 6 bestemme den stikprøvestørrelse, det vil kræve. I praksis er der naturligvis mange faktorer og overvejelser, der spiller ind på design af en given undersøgelse, fx at undersøgelser med store stikprøver normalt er forbundet med større omkostninger.

- Opgave 32** Se igen på analysen fra Epinion i eksempel 18.
- Tegn grafen for $+u$ som funktion af \hat{p} i et passende interval.
 - Ved hvilken vælgertilslutning er usikkerheden størst?
 - For hvilket parti er usikkerheden størst?
 - Hvor stor skulle stikprøven mindst have været, hvis usikkerheden for Venstre skulle være på højst 2,2% (og man dermed ikke kunne konkludere en tilbagegang med et 95% konfidensinterval)?

- Opgave 33** I 2011 viste en stor undersøgelse, at 64% af voksne danskere regelmæssigt dyrkede motion.
- I 2016 udtog man en stikprøve på 400 voksne danskere og spurgte dem, om de regelmæssigt dyrkede motion. I stikprøven var der 284, der regelmæssigt dyrkede motion.
- Vi antager i det følgende, at den sande værdi af andelen af alle voksne danskere, der i 2011 regelmæssigt dyrkede motion, var 64%.
- Bestem ud fra stikprøven i 2016 et 95% konfidensinterval for andelen af voksne danskere, der regelmæssigt dyrker motion.
 - Afgør, om andelen af voksne danskere, der regelmæssigt dyrker motion, har ændret sig.

Opgave 34 a) Bestem for $\hat{p} = 0,60$ den statistiske usikkerhed ved følgende værdier af n :

n	100	200	400	800
u				

b) Beskriv ændringen i u , når n vokser.

Opgave 35 Vis ud fra formlen i sætning 6, at man skal firedoble stikprøvestørrelsen, hvis man skal halvere usikkerheden.