MAT4010

PROSJEKTOPPGAVE:

Statistikk i $S2$

Olai Sveine Johannessen, Vegar Klem Hafnor & Torstein Mellem

20. mai 2015

Innhold

Sammendrag. Vi ser på statistikk i *Matematikk S2* og går gjennom forventningsverdi og varians, estimering, konfidensintervaller, sentralgrensesetningen og hypotesetesting med fokus på forklaringer og forståelse gjennom bruk av eksempler.

1. Stokastisk Variabel

I enkelte forsøk kan en forutsi utfallet før en faktisk har gjort forsøket. En kan for eksempel beregne hvor lang tid en kule triller ned et plan om man kjenner nødvendige parametre. I andre forsøk kan en ikke bestemme utfallet, men kun sannsynligheten for utfall. Slike forsøk kalles stokastiske forsøk og kan for eksempel være å se på summen av øynene på to terninger eller antall gevinster ved kjøp av lodd. Variabelen i kuleeksempelet kan være tid, mens variabelen i terningkastet kan være antall øyne. I stokastiske forsøk kalles variabelen en stokastisk variabel og skrives med stor bokstav, for eksempel X. Verdien til variabelen betegnes med en liten bokstav (for eksempel x). Man kan si at en stokastisk variabel er resultatet av et tilfeldig forsøk (stokastisk forsøk). Hvert mulige utfall er forbundet med en sannsynlighet for at akkurat det utfallet inntreffer. Summen av sannsynlighetene for alle mulige utfall må være 1, ettersom vi vet helt sikkert at vi får et av utfallene. Generelt kan vi si at utfallsrommet, Ω, til et stokastisk forsøk er $\{u_1, u_2, ..., u_n\}$ og de tilhørende sannsynlighetene er $\{p_1, p_2, ..., p_n\}$. Summen av sannsynlighetene er lik 1: $p_1 + p_2 + ... + p_n = 1$.

Date: 20. mai 2015.

Eksempel 1.1. Vi kjøper lodd i et lotteri hvor det blir oppgitt at hvert tredje lodd gir gevinst. Kjøper vi tre lodd har vi parametrene $p = \frac{1}{3}$ og $n = 3$, hvor p er sannsynligheten for et vellykket forsøk og n er antall slike forsøk som blir utført. Den stokastiske variabelen X velger vi til å være antall gevinster vi får av våre tre lodd.
De mulige verdiene for X blir da $0, 1, 2$ eller 3, gevinster.

For å angi sannsynligheten for et spesifikt utfall bruker vi notasjonen

$$P(X = x_i) = p_i$$

hvor x_i er et utfall og p_i dens tilhørende sannsynlighet for å inntreffe.

Eksempel 1.2. Vi fortsetter med lotteriet fra forrige eksempel og ser på sannsynlighetene for de forskjellige utfallene for den stokastiske variabelen $X =$ antall gevinster.
Sannsynlighetene er gitt ved

$$P(X = x_i) = \frac{\text{Antall gunstige}}{\text{Antall mulige}}$$

Vi ser på hvert utfall hver for seg:

$X = 0$:
Det én måte å få null gevinster på; ingen gevinst på det første loddet, ingen gevinst på det andre og ingen gevinst på det tredje. Fra kombinatorikken har vi at sannsynligheten for utfallet blir

$$P(X = 0) = \frac{2}{3} \cdot \frac{2}{3} \cdot \frac{2}{3} = \frac{8}{27}$$

$X = 1$:
Det er tre måter å få én gevinst på; enten gevinst på det første loddet, det andre loddet eller det tredje loddet. Det gir

$$P(X = 1) = 3 \cdot \left(\frac{1}{3} \cdot \frac{2}{3} \cdot \frac{2}{3} \right) = \frac{12}{27}$$

$X = 2$:
To gevinster er mulig ved å ikke få gevinst på enten det første, det andre eller det tredje loddet. Det gir

$$P(X = 2) = 3 \cdot \left(\frac{1}{3} \cdot \frac{1}{3} \cdot \frac{2}{3} \right) = \frac{6}{27}$$

$X = 3$:
Det er bare én måte å få tre gevinster på; å få gevinst på alle tre loddene. Det gir

$$P(X = 3) = \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{27}$$

Vi kan også presentere disse dataene i en tabell. I følgende vil A bety lodd som gir gevinst, mens $\bar{A}$ betyr lodd som ikke gir gevinst. Om vi tenker at

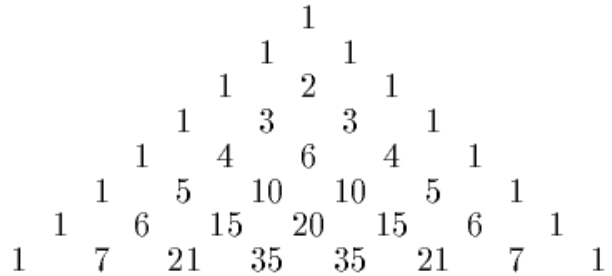
vi har 4 mulige utfall i loddeksempelet ($x_i = 0, 1, 2$ eller 3) vil utfallsrommet med tilhørende sannsynligheter kan da skrives som:

x_i	0	1	2	3
$P(X = x_i)$	$\frac{8}{27}$	$\frac{12}{27}$	$\frac{6}{27}$	$\frac{1}{27}$

For å se at dette er en fornuftig fordeling kan vi legge sammen sannsynlighetene og se at summen er 1: $\frac{8}{27} + \frac{12}{27} + \frac{6}{27} + \frac{1}{27} = \frac{27}{27} = 1$.
Vi ser at utfallsrommet består av 8 elementer om vi tar hensyn til rekkefølgen. Vi har en måte å få verdien $x_i = 0$, tre måter å få verdien $x_i = 1$, tre måter å få verdien $x_i = 2$ og en måte å få verdien $x_i = 3$:

x_i	0	1	2	3
	$\bar{A}\bar{A}\bar{A}$	$A\bar{A}\bar{A}$	$AA\bar{A}$	AAA
		$\bar{A}A\bar{A}$	$A\bar{A}A$	
		$\bar{A}\bar{A}A$	$\bar{A}AA$	

Vi ser at antall elementer fra det totale utvalgsrommet som passer til hver x_i, korresponderer til tallene i Pascals trekant. Om vi hadde kjøpt fire lodd ville vi hatt 1 måte å få null gevinster, 4 måter å få en gevinst, 6 måter å få to gevinster, 4 måter å få tre gevinster og 1 måte å få fire gevinster. Pascals trekant er dessuten en fin oversikt over binomialkoeffisientene.



FIGUR 1. Pascals trekant

1.1. **Stokastiske variable som funksjoner.** Stokastiske variable er enten diskrete eller kontinuerlige. Sannsynligheten for en gitt verdi til en kontinuerlig stokastisk variabel er lik null, mens sannsynligheten for en gitt verdi til en disktret variabel er ulik null. Når vi ser på kontinuerlige variable regner vi på sannsynligheten for at variablens verdi er i et gitt intervall. Den stokastiske variabelen i normalfordelingen er kontinuerlig, men vi kommer til å fokusere mest på disktrete variable i denne oppgaven. I eksemplene vi bruker er den stokastiske variabelen diskret.

$$\binom{0}{0}$$

$$\binom{1}{0} \quad \binom{1}{1}$$

$$\binom{2}{0} \quad \binom{2}{1} \quad \binom{2}{2}$$

$$\binom{3}{0} \quad \binom{3}{1} \quad \binom{3}{2} \quad \binom{3}{3}$$

$$\binom{4}{0} \quad \binom{4}{1} \quad \binom{4}{2} \quad \binom{4}{3} \quad \binom{4}{4}$$

$$\binom{5}{0} \quad \binom{5}{1} \quad \binom{5}{2} \quad \binom{5}{3} \quad \binom{5}{4} \quad \binom{5}{5}$$

FIGUR 2. Binomialkoeffisienter (koeffisienter fra binomialformelen $(a+b)^n$).

Det er mulig å representere både kontinuerlige og diskrete stokastiske variable som funksjoner ved å velge førsteaksen til å være mulige verdier for X, det vil si x_i, og andreaksen til å være sannsynlighetene gitt ved $P(X = x_i)$. Dette gir en sannsynlighetsfordelig til den stokastiske variabelen. En stokastisk variabel er altså en funksjon på utfallsrommet.

Vi skal nå forklare nærmere hvorfor diskrete stokastiske variable er funksjoner. En kan definere en funksjon som en liste av ordnede par. Det vi tenker på som en vanlig variabel er en liste av verdier. En kan for eksempel ha gitt variabelen tid som er definert på et intervall: $[0, 20]$ sekunder. Diskrete stokastiske variable er derimot en liste av ordnede par. Til hver mulige verdi har slike variabler en tilhørende verdi som er sannsynligheten for at verdien inntreffer.
Variabelen er gitt som en mengde mulige utfall og en tilhørende sannsynlighet. Lager man en liste av par hvor det første elementet er en mulig verdi for X og det andre elementet er utfallets sannsynlighet, $P(X = x_i)$, så har man per definisjon en funksjon på utfallsrommet.

2. Forventningsverdi

Når vi gjennomfører et stokastisk forsøk, vil den stokastiske variabelen X være resultatet av forsøket. Et spørsmål vi stiller oss før vi gjennomfører forsøket er om vi på forhånd kan vite noe om verdiene til X. Svaret er *Ja*, vi kan regne oss frem til hva vi kan forvente at X er, denne verdien kaller vi *forventningsverdien* til X vi skriver forventningsveriden til X som $E(X)$ eller μ.

Forventingsverdien er altså den verdien vi forventer at resultatet av et forsøk er, vi er ikke sikre på resultatet, men vi vet hva vi kan forvente. For å forstå hva forventningsverdi er så ser vi litt på sammenhengen mellom forventningsverdi og gjennomsnitt.

Gjennomfører vi et forsøk flere ganger, kan vi ut i fra resultatene regne ut et gjennomsnitt, dette gjennomsnittet vil da være påvirket av resultatene fra forsøket. Ser vi på forventningsverdien så er dette det matematisk korrekte svaret på hva vi kan forvente å oppnå, mens gjennomsnittet er et estimat som er tatt ut i fra et sample av forsøk. Forventningsverdien vil vi kunne regne ut før vi gjennomfører forsøket, mens gjennomsnittet er noe vi må regne ut etter forsøket, det er her viktig å få med seg at gjennomsnittet aldri vil være fullstendig lik forventningsverdien, men den vil nærme seg forventningsverdien ettersom vi gjennomfører forsøket flere ganger. Dermed vil vi kunne forutse hva gjennomsnittet vil nærme seg ved å regne ut forventningsverdien før vi gjennomfører forsøket.
Forventningsverdien til en stokastisk variabel X er regnet ut på følgende måte:

Hvis X er en stokastisk variabel som kan ha n ulike verdier da er forventningsverdien(μ) til X gitt ved følgende formel

$$\mu = E(X) = \sum_{i=1}^{n} x_i \cdot P(X = x_i)$$

Videre nå skal vi se på to eksempler hvor vi kan bruke formelen til å regne ut forventningsverdien til to forskjellige forsøk.

Eksempel 2.1. Det er et loddsalg som reklamerer med at hvert 3. lodd gir gevinst. Vi kjøper 3 lodd og regner da med å vinne på ett av loddene, stemmer dette med matematikken for loddsalget?
Vi skjønner fort at når vi kjøper 3 lodd har vi 4 mulige utfall for antall gevinster, nemlig at vi kan få 0, 1, 2 eller 3 gevinster, men hva kan vi forvente?
Her har vi satt opp en tabell over sannsynlighetsfordelingen når vi kjøper 3 lodd:

Antall gevinst x_i	Sannsynligheten for x_i $P(X = x_i)$	Sannsynlighet for å vinne $x_i \cdot P(X = x_i)$
0	$\frac{8}{27}$	$0 \cdot \frac{8}{27} = 0$
1	$\frac{12}{27}$	$1 \cdot \frac{12}{27} = \frac{12}{27}$
2	$\frac{6}{27}$	$2 \cdot \frac{6}{27} = \frac{12}{27}$
3	$\frac{1}{27}$	$3 \cdot \frac{1}{27} = \frac{3}{27}$

Setter vi nå de svarene vi har fått fra tabellen inn i formelen for å regne ut forventningsverdi får vi:

$$\mu = E(X) = 0 \cdot \frac{8}{27} + 1 \cdot \frac{12}{27} + 2 \cdot \frac{6}{27} + 3 \cdot \frac{1}{27} = \underline{1}$$

Vi ser at når vi kjøper 3 lodd kan vi forvente å vinne på ett av de.

Her ser vi at det er forskjellige sannsynligheter for å få de ulike verdien av X, men hva skjer om alle mulige utfall av X har samme sannsynlighet for å intreffe? For eksempel ved et terningkast:

Eksempel 2.2. Nå skal vi se på matematikken rundt et terningkast. Ved kast av en terning er resultatet vi kan få 1, 2, 3, 4, 5 eller 6, den stokastiske variabelen X kan altså være 1, 2, 3, 4, 5 eller 6. Men hva vil vi matematisk kunne forvente å få ved kast av én terning? Sannsynlighetsfordelingen ved kast av én terning:

Mulige x_i x_i	Sannsynligheten for x_i $P(X = x_i)$	Sannsynlighet for å vinne $x_i \cdot P(X = x_i)$
1	$\frac{1}{6}$	$1 \cdot \frac{1}{6} = \frac{1}{6}$
2	$\frac{1}{6}$	$2 \cdot \frac{1}{6} = \frac{2}{6}$
3	$\frac{1}{6}$	$3 \cdot \frac{1}{6} = \frac{3}{6}$
4	$\frac{1}{6}$	$4 \cdot \frac{1}{6} = \frac{4}{6}$
5	$\frac{1}{6}$	$5 \cdot \frac{1}{6} = \frac{5}{6}$
6	$\frac{1}{6}$	$6 \cdot \frac{1}{6} = \frac{6}{6}$

Setter vi nå disse verdiene inn i formelen for utregning av forventningsverdi får vi:

$$\mu = E(X) = \frac{1}{6} + \frac{2}{6} + \frac{3}{6} + \frac{4}{6} + \frac{5}{6} + \frac{6}{6} = \underline{\underline{3,5}}$$

Vi ser altså at når vi kaster en terning, kan vi forvente at $X = 3,5$.

Dette ser vi at aldri vil stemme, men det er fortsatt matematisk korrekt.

Ser vi på det siste eksempelet en gang til så er ikke forventningsverdien på $3,5$ så dum. Som nevnt tidligere så er forventningsverdien og gjennomsnittetto verdier som vil være tilnærmet like. Dette vil da si at dersom vi kaster en terning mange ganger, for deretter å summere alle verdiene vi får og dele på antall kast, altså finne gjennomsnittet, så ser vi at dette gjennomsnittet vil nærme seg $3,5$ som er forventningsverdien. Gjennomsnittet vil aldri kunne bli helt lik $3,5$ den vil bare være tilnærmet lik.

3. Varians og standardavvik

Dersom vi gjenomfører et stokastisk forsøk mange ganger vil den stokastiske variabelen X få mange forskjellige verdier. Forventningsverdien til X vil da fortelle oss hva gjennomsnittet av alle verdiene av X er. Det vi ikke vet er hvordan verdiene til X vil variere fra forsøk til forsøk, variasjonen i verdiene til X kaller vi for spredningen til X. Et mål på denne spredningen er variansen til X og vi skriver $\text{Var}(X)$. Variansen er altså et spredningsmål på sannsynlighetsfordlingen til X, når vi måler denne spredningen på verdiene til X tar vi utgangspunkt i forventningsverdien og ser på hvor mye hver av verdiene avviker fra forventningsverdien. Formelen for varians er gitt slik:

Variansen til en stokastisk variabel X med forventningsverdi μ er gitt ved

$$\text{Var}(X) = \sum_{i=1}^{n} (x_i - \mu)^2 \cdot P(X = x_i)$$

Her er da alle $x_i = x_1, x_2, ...x_n$ mulige verdier av X

Vi ser at variansen er et mål på forskjellen mellom μ og x_i vektet med hvor sannsynlig det er for at x_i intreffer.
Et lite problem med å sammenligne variansen($\text{Var}(X)$) med forventningsverdien(μ) er at de ikke har samme benevning. Dersom μ er målt i meter(m) så vil $\text{Var}(X)$ være målt i kvadratmeter(m^2). Hvis vi ønsker at spredningsmålet har samme benevning som forventningsverdien, tar vi kvadratroten av variansen og finner *standardavviket* som vi skriver $SD(X)$ eller σ:

Standardavviket for en stokastisk variabel X med forventningsverdi μ og variansen $\text{Var}(X)$ er gitt ved

$$\sigma = SD(X) = \sqrt{\text{Var}(X)}.$$

Varians og standardavvik er mål på hvor spredd resultatene fra et forsøk er sammenlignet med forventningsverdien μ.

Vi ser på eksempel 2.1 igjen, men denne gangen skal vi konsentrere oss om å variansen og standardavviket til dette forsøket, hvor mye vil verdiene våre variere fra forventningsverdien?

Eksempel 3.1. Vi kjøper fortsatt 3 lodd og statistisk sett skal hvert 3. lodd gi gevinst. Vi har allerede funnet ut i eksempel 2.1 at forventningsverdien er lik 1 ($\mu = 1$), men hva er variansen og standardavviket?
Vi setter opp i en tabell for å få oversikt:

Antall x_i	Sannsynlighet $P(X = x_i)$	P(gevinst) $x_i \cdot P(X = x_i)$	Vektet sannsynlighet $(x_i - \mu)^2 \cdot P(X = x_i)$
0	$\frac{8}{27}$	$0 \cdot \frac{8}{27} = 0$	$(0-1)^2 \cdot \frac{8}{27} = \frac{8}{27}$
1	$\frac{12}{27}$	$1 \cdot \frac{12}{27} = \frac{12}{27}$	$(1-1)^2 \cdot \frac{12}{27} = 0$
2	$\frac{6}{27}$	$2 \cdot \frac{6}{27} = \frac{12}{27}$	$(2-1)^2 \cdot \frac{6}{27} = \frac{6}{27}$
3	$\frac{1}{27}$	$3 \cdot \frac{1}{27} = \frac{3}{27}$	$(3-1)^2 \cdot \frac{1}{27} = \frac{4}{27}$

Vi setter veridene fra tabellen inn i formelen for variansen og får:

$$\begin{aligned} \mathrm{Var}(X) &= \frac{8}{27} + 0 + \frac{6}{27} + \frac{4}{27} \\ \mathrm{Var}(X) &= \frac{18}{27} \\ \mathrm{Var}(X) &= \underline{\underline{\frac{2}{3}}} \end{aligned}$$

Variansen er altså $\frac{2}{3}$.
Standardavviket blir da

$$\begin{aligned} \sigma &= \sqrt{\mathrm{Var}(X)} = \sqrt{\frac{2}{3}} \\ \sigma &= \underline{\underline{0,816}} \end{aligned}$$

Vi ser har at når vi kjøper 3 lodd og forventer å vinne på ett av de tre loddene, vil standardavviket være $0,816$

4. Binomisk fordeling

Før vi går videre og ser på normalfordelingen så må vi forklare hva et binomisk forsøk er og hvordan vi kan regne med en binomisk fordeling.

Et binomisk forsøk består av n uavhengige delforsøk, hvor sannsynligheten for at en hendelse A intreffer er lik p i hvert av delforsøkene. I et slikt forsøk vil X være antallet ganger A intreffer. Da vet vi at forventningsverdien $E(X)$ og variansen $\mathrm{Var}(X)$ til X er gitt ved:

$$\begin{aligned} E(X) &= n \cdot p \\ \mathrm{Var}(X) &= n \cdot p \cdot (1-p) \end{aligned}$$

Ser vi tilbake på lodd-eksempelet (eksempel 2.1) vi brukte tidligere ser vi nå at dette faktisk er en binomisk fordeling, hvor $p = \frac{1}{3}$ og hendelse $A =$ Loddet gir gevinst.

5. Normalfordelingen

Når vi ser på sannsynligheter for å få gevinst i et forsøk, eller andre forsøk med sannsynligheter vil vi alltid kunne skrive opp resultatene i en sannsynlighetsfordeling. Forskjellige forsøk vil naturlignok ha forskjellige sannsynlighetsfordelinger. Det er slik at normalfordelingen blir sett på som den viktigste av fordelingene. Det vil være enkelte sannsynlighetsfordelinger som vil bli tilnærmet lik normalfordelingen dersom vi øker antallet ganger vi gjennomfører forsøket, dette skal vi se på litt senere.

Det var den tyske matematikeren Carl Friedrich Gauss (1777 - 1855) som først kom med denne fordelingen og derfor kalles den også Gauss-fordelingen. Først og fremst er det viktig å få med seg at normalfordelingen kun vil være en tilnærming til den faktiske situasjonen. Normalfordelingen beskriver variasjonen i visse typer data i tillegg til at den er en god tilnærming til andre fordelinger. Sannsynlighetsfordelingen til et forsøk kan man sette inn i et koordinatsystem, da er det letteste å se resultatet i et histogram. Ser vi på tetthetsfunksjonen til dette histogrammet vil det bli mer og mer lik en normalfordelingsfunksjon når vi gjentar forsøket flere ganger, altså øker variabelen n.

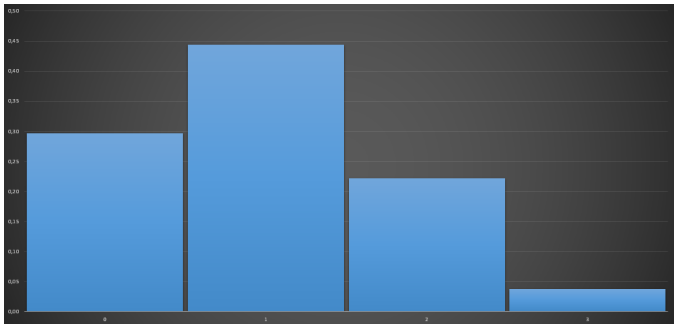
Vi skal nå se på sannsynlighetsfordelingen til lodd-forsøket fra eksempel 2.1. Denne sannsynlighetsfordelingen vil være binomisk fordelt og vi skal se på hva som skjer når vi øker n, altså antall lodd vi kjøper, eller sett på en annen måte antall ganger vi gjennomfører forsøket.

Ser vi på sannsynlighetsfordelingen til X når vi kjøper 3 lodd ($n = 3$), vil den se slik ut(lik som eksempel 2.1):

Antall gevinst x_i	Sannsynligheten for x_i $P(X = x_i)$	Sannsynlighet for å vinne $x_i \cdot P(X = x_i)$
0	$\frac{8}{27}$	$0 \cdot \frac{8}{27} = 0$
1	$\frac{12}{27}$	$1 \cdot \frac{12}{27} = \frac{12}{27}$
2	$\frac{6}{27}$	$2 \cdot \frac{6}{27} = \frac{12}{27}$
3	$\frac{1}{27}$	$3 \cdot \frac{1}{27} = \frac{3}{27}$

Setter vi verdiene fra kolonne en og to inn i en graf får vi følgende histogram over sannsynlighetsfordelingen.

Dette stemmer bra med eksempel 2.1 hvor vi fant ut at det er mest sannsynlig og vi kan forvente å vinne på ett lodd når vi kjøper tre. Vi ser av grafen at det er søylen når $X = 1$ som er høyest, som betyr at dette er mest

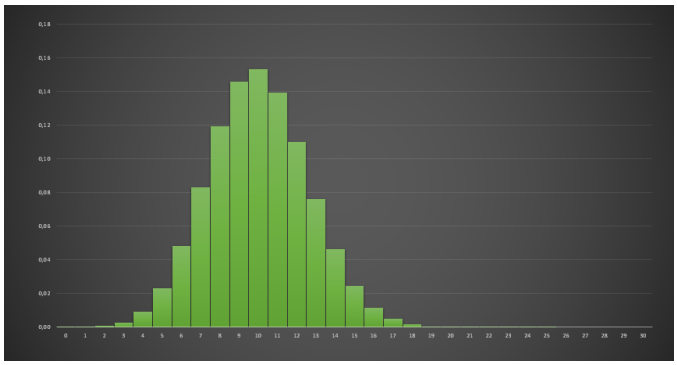


FIGUR 3. Sannsynlighetsfordeling når vi kjøper 3 lodd ($n = 3$)

sannsynlig. Ut ifra histogrammet kan vi lese av hvor stor sannsynlighet det er for at hver av de forskjellige verdiene av X intreffer. Dette vil være høyden til den bestemte verdien, fordi alle søylene har bredde lik 1.

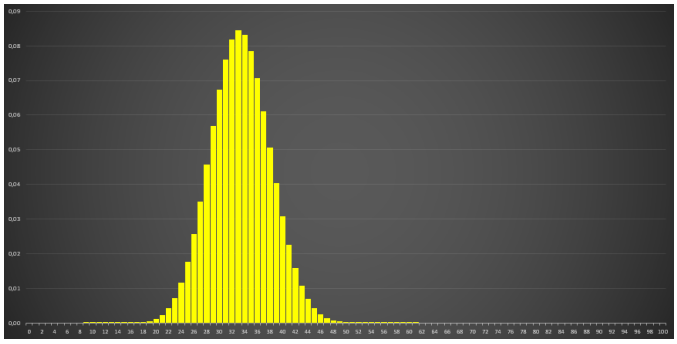
Dersom vi kjøper flere enn tre lodd, vil jo naturlig nok forventningsverdien endre seg, men hvordan endrer sannsynlighetsfordelingen seg? Dersom vi for eksempel kjøper 30 lodd i stedet for 3, da vil forventningsverdien være lik 10, $\mu = 10$. Men hvordan ville variansen og standardavviket bli og hvordan vil histogrammet som viser sannsynlighetsfordelingen se ut? Altså hvoran ser sannsynlighetene for å vinne for eksempel 11, 12, 9 eller 21 lodd når vi kjøper 30, det er dette histogrammet over sannsynlighetsfordelingen forteller oss.

Dersom vi hadde kjøpt 30 lodd, ville sannsynlighetsfordelingen til X se slik ut:



FIGUR 4. Sannsynlighetsfordeling når vi kjøper 30 lodd ($n = 30$)

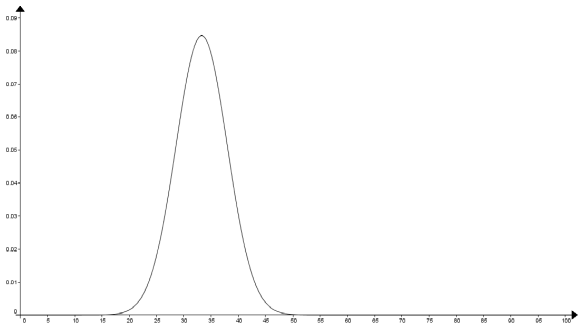
Vi ser her når vi kjøper 30 lodd vil det være størst sannsynlighet for at vi vinner på 10 av disse, vi kan også her lese av sannsynlighetene for de forskjellige verdiene av X på andreaksen.
Til slutt ser vi på hvordan sannsynlighetsfordelingen ser ut når vi tenker oss at vi kjøper 100 lodd



FIGUR 5. Sannsynlighetsfordeling når vi kjøper 100 lodd ($n = 100$)

Her ser vi at forventningsverdien ligger rundt 33, fordi det er når $X = 33$ at vi får den høyeste stolpen. Vi kan også her lese av hvor sannsynlig det er for at en bestemt X intreffer ved å se på y-verdien til denne X'en.

Vi ser at formen til histogrammet blir mer og mer jevn etterhvert som vi øker n, den går mot det vi kaller en klokkeform. Vi ser at for en binomisk fordeling vil det være slik at når vi øker antall ganger vi gjennomfører forsøket, så vil fordelingen bli tilnærmet normalfordelt. Det var nettopp dette Gauss også kom frem til, han kom da også frem til normalfordelingsfunksjonen som ser slik ut:



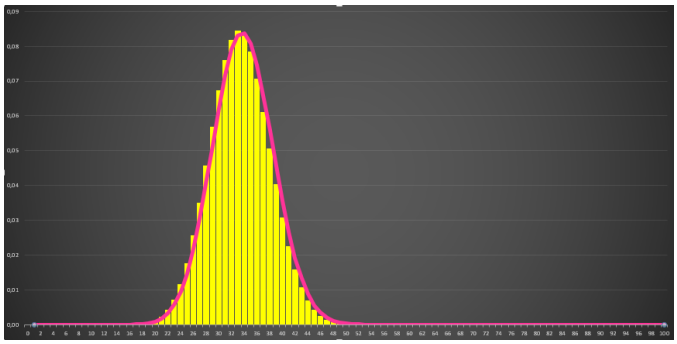
FIGUR 6. Normalfordelingskurve når $\mu = 33,33$ og $\sigma = 4,71$

Normalfordelingsfunksjonen vil ha varianter av denne formen for alle verdier av μ og σ.

Hvis X er en stokastisk variabel med forventningsverdi μ og standardavvik σ så sier vi at X er tilnærmet normalfordelt dersom histogrammet til sannsynlighetsfordelingen faller omtrent sammen med grafen til normalfordelingsfunksjonen.

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Ser vi på hvordan vårt histogram passer med normalfordelingsfunksjonen når $n = 100$ får vi en slik graf: Vi ser at normalfordelingsfunksjonen og



FIGUR 7. Histogram når $n = 100$ og grafen til normalfordelingsfunksjonen når $\mu = 33,33$ og $\sigma = 4,71$

histogrammet nesten sammenfaller og derfor sier vi at den binomisk fordelte variabelen X er tilnærmet normalfordelt.

6. STATISTIKK

Eksempler på statistiske data kan være meningsmålinger, data om temperaturendringer eller data fra eindomsmarkedet. Generelt kan vi si at man samler inn informasjon eller data for så å bruke dette til å si noe om de faktiske forhold. I statistikk bruker en sannsynlighetsmodeller til å tolke og analysere virkeligheten. Her bruker vi matematikken som et verktøy til å analysere statistiske data. Vi ser nærmere et eksempel på en meningsmåling.

Eksempel 6.1. Vi spør 500 tilfeldige mennesker hvilket parti de ville stemt på om valget var i dag. 75 stk sa de ville stemt Høyre.

Når en ønsker å finne ut av hvor mange som stemmer Høyre, er det umulig å spørre alle Norges innbyggere med stemmerett. Man må spørre et utvalg av mennesker. Deretter kan man tolke og analysere disse resultatene med hensikt i å få et fornuftig bilde av virkeligheten. Vi ønsker å finne ut av hva som er tilfeldigheter og hva som er faktiske trender. Det kan være nyttig å ha i bakhodet at statistikk i seg selv ikke er matematikk, men en bruker matematiske modeller når vi skal analysere og vurdere i statistikken.

6.1. **Estimering.** Siden vi ikke kjenner de faktiske forhold, lager vi estimater (tilnærminger). I meningsmålingseksempelet vil $\hat{p} = \frac{X}{n} = \frac{75}{500} = 0.15$ være et estimat for sannsynligheten for at en tilfeldig person med stemmerett i Norge stemmer Høyre. Vi kjenner ikke den faktiske sannsynligheten, p, for at en person stemmer Høyre, men anslår en verdi ut i fra dataene våre.

Estimatet $\hat{p} = 0.15$ kalles et punktestimat. Vi skal være ganske heldige om den virkelige sannsynligheten er akkurat $p = 0.15$. For å bestemme hvor det er sannsynlig å finne p kan vi lage et intervallestimat for $\hat{p}$ (altså et område vi mener det er sannsynlig å finne p). Et slikt intervall kalles konfidensintervall.

Siden 500 er et lite tall i forhold til innbyggertallet vil ikke det at vi trekker ut en person av befolkningen forandre sannsynligheten for om neste person er Høyrevelger eller ikke. Det at vi trekker ut en person av befolkningen forandrer ikke sannsynlighetsfordelingen til den resterende befolkningen (Dette kan vi fint gjøre 500 ganger uten at det forandrer sannsynlighetsfordelingen). Delforsøkene er uavhengige og sannsynlighetsfordelingen til forsøket er dermed binomisk. Vi har at $\mu = np$ og $\sigma^2 = np(1-p)$.

Når utvalget blir stort nok vil den binomiske sannsynlighetsfordelingen gå mot å bli normalfordelt. En regel man bruker for å vurdere om utvalget er stort nok er: $(n\hat{p} > 5)$. I dette eksempelet ser vi at vi er innenfor dette kravet: $(n\hat{p} = 500 \cdot 0.15 = 75 > 5)$. For alle normalfordelinger er sannsynligheten for at variabelens verdi er innenfor $\pm\, 1,96\sigma$ ca. 95%.

En god estimator $\hat{p}$ vil være slik at gjennomsnittet av flere $\hat{p}$, vil nærme seg p. Om dette er tilfellet sier vi at estimatoren er forventningsrett: $E(\hat{p}) = p$. Gjennomsnittet til $\hat{p}$ vil ha en sannsynlighetsfordeling med p som forventningsverdi. Dessuten må denne fordelingsfunksjonen bli høyere og smalere (mer konsentrert om p) når n øker. Det vil si at $\mathrm{Var}(\hat{p})$ bør bli mindre og mindre nå n øker. Vi ser at dette er tilfellet for estimatoren $\hat{p}$ i meningsmålingseksempelet:

$$E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}np = p$$
$$\mathrm{Var}(\hat{p}) = Var\left(\frac{X}{n}\right) = \frac{1}{n^2}\mathrm{Var}(X) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

Et 95% konfidensintervall for $\hat{p}$ er intervallet $[\hat{p} - 1.96\hat{\sigma}\ ,\ \hat{p} + 1.96\hat{\sigma}]$, der $\hat{\sigma}$ er en estimator for standardavviket til $\hat{p}$. Vi sier at dette estimatet har sikkerhet på 95%. For å lage et godt estimat for p, bør vi regne ut flere konfidensintervall (altså gjøre flere meningsmålinger). 95% av disse konfidensintervallene vil da inneholde verdien p. Nedenfor skal vi vise hvordan en beregner et konfidensintervall.

Vi bruker variabelen $Z = \frac{X-np}{\sqrt{np(1-p)}}$ som er variabelen for standardnormalfordelingen. $X = n\hat{p}$ er antall Høyrevelgere i utvalget vårt.

$$P\left(-1.96 \leq \frac{X-np}{\sqrt{np(1-p)}} \leq 1.96\right) \approx 0.95$$

Trekker ut n i teller og nevner og stryker mot hverandre:

$$P\left(-1.96 \leq \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq 1.96\right) \approx 0.95$$

Siden vi ikke kjenner p, erstatter vi p med $\hat{p}$ i nevner. Dette er ikke helt riktig, her kommer det egentlig inn teori som vi ikke skal gå inn på i denne oppgaven. Vi erstatter derfor p med $\hat{p}$ her, siden dette nesten er riktig:

$$P\left(-1.96 \leq \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \leq 1.96\right) \approx 0.95$$

Dette gir

$$P\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \approx 0.95$$

Dermed har vi et 95% konfidensintervall for p:

$$\left[\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ,\ \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Innfører symbolet $\hat{\sigma}$ for standardavviket til $\hat{p}$:

$$[\hat{p} - 1.96\hat{\sigma}\ ,\ \hat{p} + 1.96\hat{\sigma}]$$

7. Sentralgrensesetningen

I 1733, altså lenge før Gauss studerte normalfordelingen fant Abraham de Moivre ut at det gikk an å finne en tilnærming til binomiske sannsynligheter ved å bruke standardnormalfordelingen. Senere ble det oppdaget at den tilnærmingen de Moivre brukte var et spesialtilfelle som senere er blitt kjent som *sentralgrensesetningen.*

Sentralgrensesetningen

La X være en uavhengig stokastisk variabel med forventningsverdi μ og standardavvik σ. La da S være summen av n verdier av X, altså $S = X_1 + X_2 + ... + X_n$. Da er S tilnærmet normalfordelt og har forventningsverdi $E(S) = n \cdot \mu$ og standardavvik $SD(S) = \sqrt{n} \cdot \sigma$

Tilnærmingen er best for store verdier av n.

Det vi ser fra setningen over er at alle X_n har samme sannsynlighetsfordeling, men det er ikke stilt noen krav til hvilken type fordeling de skal ha, sentralgrensesetningen gjelder uansett om X_n er kontinuerlig eller diskret. Når vi tidligere så på normalfordelingen så var det slik at enkelte andre fordelinger vil bli tilnærmet like normalfordelingen for store n, men enkelte

fordelinger vil aldri bli tilnærmet normalfordelt. Det da sentralgrensesetningen sier er at dersom vi tar summen av noen stokastiske variable i en hvilken som helst fordeling, så vil summen være tilnærmet normalfordelt. Grunnen til dette ligger i beviset, men vi kommer ikke inn på dette beviset i denne oppgaven. Vi skal videre forklare litt om hvordan sentralgrensesetningen brukes ved hjelp av et par eksempler.

Noe av det som gjør sentralgrensesetningen så interessant er at den gir oss mulighet til å regne ut sannsynligheter i situasjoner der det ellers ikke hadde vært mulig, slik som vi skal se i eksempelet under.

Eksempel 7.1. Når vi kaster en terning, har vi den stokastiske variabelen $X = \textit{antall øyne}$, utfallsrommet til X vil være $(1, 2, 3, 4, 5, 6)$. Dermed ser vi at X er uniformt fordelt, som betyr at vi vil i lengden få tilnærmet like mange av hvert antall øyne. Uansett hvor mange ganger vi kaster terningen så vil vi ikke kunne si at X blir tilnærmet normalfordelt. Forventnignsverdien når vi kaster en terning vet vi er $E(X) = \mu = 3,5$ og standardavviket er $SD(X) = \sqrt{\frac{35}{12}}$.

Eksempel 7.2. Vi tenker oss nå at vi kaster 1000 terninger. Vi lar S være summen av antall øyne til de 1000 terningene. Som sagt er forventningsverdien og standardavviket for antall øyne på hver av terningene lik $E(X) = \mu = 3,5$ og $SD(X) = \sigma = \sqrt{\frac{35}{12}}$
Vi finner forventningsverdien og standardavviket til S ved å bruke sentralgrensesetningen:

$$E(S) = \mu \cdot n = 3,5 \cdot 1000 = \underline{\underline{3500}} \tag{7.1}$$

$$SD(S) = \sigma\sqrt{n} = \sqrt{\frac{35}{12}} \cdot \sqrt{1000} = \underline{\underline{54,0}} \tag{7.2}$$

Grunnen til at denne summen da vil være normalfordelt er fordi vi tenker oss at summen av de 1000 terningene ikke vil være 3500 hver gang, de 1000 terningene vil ha en sum som ligger i området rundt 3500. Denne summen vil være normalfordelt med 3500 som forventningsverdi og et standardavvik på 54.

Fra eksemplet over ser vi at utregning av sannsynligheter for eksempel med kast av > 1000 terninger er mulig fordi vi ved hjelp av sentralgrensesetningen finner forventningsverdi og standardavvik, dermed kan vi bruke standardnormalfordelingen og finne sannsynlighetene vi trenger.

Som en liten oppsummering kan vi si at sentralgrensesetningen handler om at summen av noen stokastiske variablene er normalfordelt uavhengig av hvilken fordeling de stokastiske variable har. Dermed åpner det seg en helt ny måte å regne sannsynligheter for vanskelige fordelinger på.

8. Hypotesetesting

En hypotese er et utsagn om eller en påstand om en ukjent parameter i en statistisk modell. Et slikt utsagn kan for eksempel være på formen " det er 30% sannsynlighet for å vinne på lodd" eller "minst 20% av loddene gir gevinst".
Hypotesetesting tar utgangspunkt i et utvalg av stokastiske forsøk for å si noe om hvor sikkert vi kan komme med vårt utsagn.

Eksempel 8.1. I et Ludospill bruker spillerne hver sin terning. Ole synes at Ivar har fått påfallende mange seksere i løpet av spillet. For å se om Ivar jukser bestemmer Ole seg for å registrere antall seksere Ivar kaster i løpet av sine neste 24 kast.

Dersom Ivar ikke jukser vil sannsynligheten for at han kaster en sekser være $p = \frac{1}{6}$. Dersom han jukser på en måte som gjør at han kaster flere seksere vil $p > \frac{1}{6}$.

Påstanden Ole vil teste er om $p > \frac{1}{6}$. Men hvor mange seksere må Ivar kaste før Ole med sikkerhet kan beskylde ham for juks? Det er forventet at Ivar kaster $24 \cdot \frac{1}{6} = 4$ seksere, men siden terningkast er et stokastisk forsøk kan det tenkes at han får fem eller seks seksere uten at han jukser. Ole bestemmer seg for at han vil være 95% sikker på at Ivar jukser før han sier noe. Det vil si at han må finne ut hvor mange seksere det er som til sammen har 5% sannsynlighet for å inntreffe.

Innfører vi statistisk notasjon i Oles problemstilling ønsker han å finne når

$$P\left(\text{Påstå at Ivar jukser}|\text{Ivar ikke jukser}\right) \leq 0.05$$

Betingelsen "Påstå at Ivar jukser" blir den ukjente variabelen i denne problemstillingen og for å finne den må vi se på den kumulative sannsynlighetsfordelingen for forsøket.

x	4	5	6	7	8
$P(X = x)$	0.2139	0.1711	0.1084	0.0557	0.0237
$P(X \geq x)$	0.5885	0.3704	0.1995	0.0912	0.0354

Hvor vi bruker at $P(X \geq x) = 1 - P(X \leq x - 1)$ ettersom terningkast har diskrete utfall.

Fra tabellen ser vi nå at det er 3.54% sannsynlig at Ivar kaster åtte eller flere seksere. Så dersom Ole påstår at Ivar jukser om Ivar får åtte eller flere seksere er det under 5% sannsynlighet for at han beskylder Ole for noe han ikke har gjort.

8.1. **Gangen i hypotesetesting.** I hypotesetesting går man gjennom fire punkter og vi skal knytte disse opp til det foregående eksempelet:

(1) Formulere problemet ved hjelp av *nullhypotese*, H_0, og *alternativ hypotese*, H_1.
(2) Bestemme signifikansnivået.
(3) Gjennomføre testen.
(4) Konkludere.

8.1.1. *Formulere problemet.* Når Ole ville beskylde Ivar for juks var hans påstand at $p > \frac{1}{6}$ for at Ivar kastet en sekser. Den opprinnelige antagelsen, som begge spilte under til å begynne med var at $p = \frac{1}{6}$. Utgangspunktet for testen kalles *nullhypotesen* og betegnes H_0. I dette tilfellet blir da

$$H_0 : \quad p = \frac{1}{6}$$

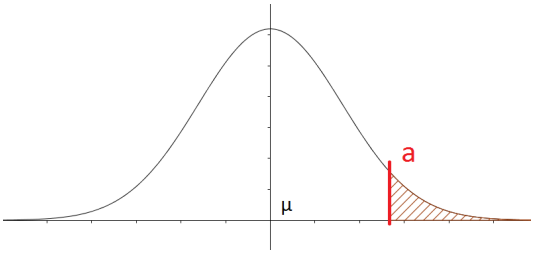
Påstanden som skal prøves ut kalles den *alternative hypotesen* eller mothypotesen, H_1:

$$H_1 : \quad p > \frac{1}{6}$$

Vi sier at vi vil teste

$$H_0 : \ p = \frac{1}{6} \quad \text{mot} \quad H_1 : \ p > \frac{1}{6}$$

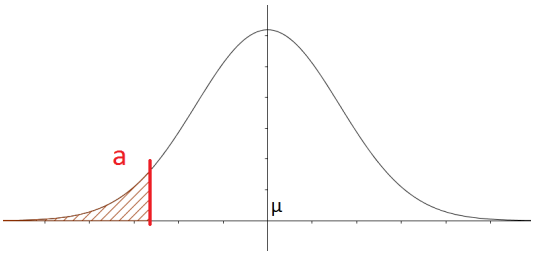
Denne typen test kalles en *høyresidig* test, ettersom vi ønsker å minimere muligheten til å ta feil i det øvre sjiktet av sannsynlighetsfordelingen.



En *venstresidig* test vil være på formen

$$H_0 : \ p = p_0 \quad \text{mot} \quad H_1 : \ p < p_1$$

og vi ønsker å minimere muligheten for å ta feil i det nedre sjiktet.



Den siste mulige formuleringen for en hypotesetest kalles en *to-sidig* test:

$$H_0 : \ p = p_0 \quad \text{mot} \quad H_1 : \ p \neq p_1$$

og man ønsker å minimere muligheten for å komme med en feilaktig påstand i begge ender av sannsynlighetsfordelingen.

8.1.2. *Bestemme signifikansnivået.* Når vi gjennomfører en hypotesetest ønsker vi å sette et nivå for hvor stor muligheten for å komme med en feilaktig påstand skal være. Dette nivået kalles signifikansnivået. Om vi som i eksempelet over ønsker å ha en mindre enn 5% sannsynlighet for å påstå at Ivar jukser når han i virkeligheten ikke gjør jukser, har vi satt signifikansnivået til å være 0.05.

Signifikansnivået er hvor liten sannsynlighet vi ønsker for å komme med en feilaktig påstand.

8.1.3. *Gjennomføre testen.* Avhengig av om man har valgt en høyresidig eller venstresidig test ønsker man å finne for hvilke verdier av den stokastiske variable vi kan komme med en påstand som har lavere sannsynlighet enn signifikansnivået for å være falsk.

I Oles tilfelle påstår han at sannsynligheten for at Ivar kaster en sekser er større enn den skal være, det vil si at hans mothypotese er at $p \geq p_0$. Dermed må Ole gjennomføre en høyresidig test.

For en *høyresidig* test får problemet formen:

$$P\left(H_1|H_0\right) \geq s$$

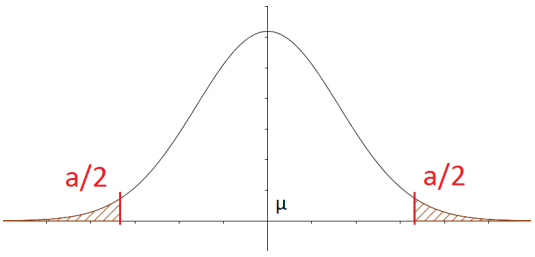
Vi velger så en testoperator X og tar en stikkprøve der X får verdien a. Vi regner så ut $P(X \geq \mathrm{a})$.

For en *venstresidig* test:

$$P\left(H_1|H_0\right) \leq s$$

Vi velger så en testoperator X og tar en stikkprøve der X får verdien a. Vi regner så ut $P(X \leq \mathrm{a})$.

Skal man å gjøre en *to-sidig* test velger vi en testoperator X og tar en stikkprøve der X får verdien a. Verdiene vi setter inn blir da $\frac{\mathrm{a}}{2}$. Vi regner så ut $P\left(X \leq \frac{\mathrm{a}}{2}\right)$ og $P(X \geq \frac{\mathrm{a}}{2})$.



8.1.4. *Konkludere.* Etter at man har regnet ut den samsvarende sannsynligheten for testen, denne kalles P-verdien, sammenligner man den med signifikansnivået. Dersom P-verdien er lavere enn signifikansnivået forkaster man *nullhypotesen*, H_0 og påstår *mothypotesen*, H_1.

For en *to-sidig* test må summen av P-verdiene være mindre enn signifikansnivået.

8.2. **Hypotesetesting av forventningsverdier.** Når man skal gjøre hypotesetester av forventningsverdier gjør man i prinsipp det samme som i en vanlig hypotesetest. Når vi vil komme med påstander om populasjonsforventningen, μ, tar vi utgangspunkt i estimatoren $\hat{X}$. Estimatoren er gjennomsnittet av verdiene fra n utvalg. Vi kan anta at estimatoren er normalfordelt, enten fordi det stokastiske forsøket er normalfordelt eller fordi vi gjennom sentralgrensetningen har at $\hat{X}$ er normalfordelt ved passende stor n.

Vi bruker eksempelet med Ole og Ivar som spiller ludo igjen for å se hvordan man kan teste påstander om forventningsverdien.

Eksempel 8.2. Resultatet av et terningkast er normalfordelt med forventningsverdien $\mu_0 = 4$. Standardavviket vil være $\sigma_0 = \sqrt{\frac{10}{3}} = 1.8257$.

Ivar kastet syv seksere. Ole lurer på om dette er langt over normen eller innenfor variasjon som følge av tilfeldigheter. Ole bestemmer seg for at dersom det er 5%, eller mindre, sannsynlig for å få syv seksere, så vil han påstå at forventningsverdien for antall seksere er høyere enn 4.

Oles hypotesetest blir da

$$H_0: \ \mu = 4 \quad \text{mot} \quad H_1: \ \mu \geq 4$$

med et signifikansnivå på 5%.

Ole regner ut P-verdien gitt ved

$$P(\hat{X} \geq 7 | \mu = 4)$$

Fra tabellen i forrige eksempel har vi at $P = 9.12\%$. Ettersom Ole satte signifikansnivået for testen til å være 5% beholder han *nullhypotesen.*