Sandsynlighedsregning og statistik med binomialfordelingen

Katja Kofod Svan og Olav Lyndrup

Januar 2019

Indhold

Stokastiske variable 3
Middelværdi og spredning 6
Binomialfordelingen 11
Andre sandsynlighedsfordelinger 19
Normalfordelingen 19
Binomialtest 22
Konfidensintervaller 29
Appendix 35
1. Hypergeometrisk fordeling 35
2. Bevis for middelværdi om binomialfordeling 37

Stokastiske variable

Eksperimenter, der ikke kan forudsiges, kaldes *stokastiske eksperimenter*. De variable man arbejder med i stokastiske eksperimenter kaldes *stokastiske variable*.

Stokastiske variable optræder blandt andet i beskrivelse af en række spilsituationer, som terningkast, lottospil, kortspil mm., men også i flere andre situationer som for eksempel ved beregning af sandsynligheden for at få fem børn af samme køn.

Til et stokastisk eksperiment hører et udfaldsrum $U = \{u_1, u_2, ..., u_n\}$, som består af *mængden af* de mulige udfald. For eksempel vil udfaldsrummet ved kast med én terning, hvor man tæller øjne, være $U = \{1,2,3,4,5,6\}$.

Traditionelt benyttes stort U til at betegne udfaldsrummet, mens lille u benyttes til at betegne et enkelt udfald i udfaldsrummet. De enkelte udfald nummereres ved hjælp af et index i, som kan antage værdierne fra 1 til n, når der er n udfald i udfaldsrummet.

Til hvert af udfaldene u_i i udfaldsrummet U knyttes en sandsynlighed p_i, som er et tal mellem 0 og 1.

I eksemplet med terningen er sandsynligheden for et bestemt antal øjne, dvs. et af de seks udfald $u_1, u_2, ..., u_6$, den samme, nemlig $\frac{1}{6}$. Bogstavet p kommer fra engelsk "probability".

Bemærk, at der benyttes samme index på sandsynlighederne p som på udfaldene u, dvs. til udfaldet u_i hører sandsynligheden p_i.

Sammenhængen mellem udfald og sandsynlighed opskriver vi ofte i en sandsynlighedstabel:

Udfald u	u_1	u_2	…	…	u_n
Sandsynlighed p	p_1	p_2	…	…	p_n

Summen af alle sandsynlighederne vil altid være 1.

Hvis vi for eksempel kun er interesserede i nogle af udfaldene i udfaldsrummet, så samler vi dem i en delmængde af U og kalder denne delmængde for *en hændelse*, som vi betegner H.

Vi sammenfatter ovenstående i en definition:

Definition 1 **Sandsynlighedsfelt**

Et endeligt udfaldsrum U med tilhørende sandsynlighedstabel kaldes for et *endeligt sandsynlighedsfelt*.

De enkelte sandsynligheder ligger mellem 0 og 1: $0 \le p_i \le 1, \quad i = 1...n$.

Summen af sandsynlighederne giver 1: $p_1 + p_2 + ... + p_n = 1$.

En delmængde H af udfaldsrummet kaldes en *hændelse*. Sandsynligheden for hændelsen H er summen af de sandsynligheder, der hører til de enkelte udfald i hændelsen.

Hvis alle udfald i U har samme sandsynlighed, så kaldes sandsynlighedsfeltet for et *symmetrisk sandsynlighedsfelt*.

Eksempel 1 **Kast med en terning**

Hvis vi kaster en terning og lader den stokastiske variabel X tælle antallet af øjne på terningen, så har vi et symmetrisk sandsynlighedsfelt med følgende sandsynlighedstabel:

Stokastisk variabel $X = x_i$	1	2	3	4	5	6
Sandsynlighed $P(X = x_i)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Af denne tabel kan man for eksempel aflæse, at der er $\frac{1}{6}$ sandsynlighed for at få en 2'er, hvilket også kan skrives som $P(X = 2) = \frac{1}{6}$.

At slå en 1'er eller en 5'er i et kast med en terning er et eksempel på en hændelse, $H = \{1,5\}$. Vi siger, at 1'ere eller 5'ere er gunstige udfald for undersøgelse af hændelsen H, fordi det netop er disse de udfald, der indgår i hændelsen. I hændelsen H ser vi altså bort fra de andre udfald 2, 3, 4 og 6.

Eksempel 2 **Det tyske Lotto**

Grafik: www.colourbox.dk

I det tyske Lotto er der 49 kugler i en beholder, og kuglerne er nummereret med tallene $1, 2, \ldots, 49$. Fra beholderen trækkes én kugle ad gangen, og der trækkes seks kugler i alt som *tilsammen* udgør et udfald. Når vi 'trækker' et udfald bestående af 6 kugler, kan man fx tænke på situationen, som en tabel med 6 celler, der skal udfyldes med de tal, vi trækker. Første gang vi trækker en kugle, er der altså 49 kugler, dvs. tal, at vælge imellem. Anden gang er der så 48 muligheder osv.

49	48	47	46	45	44

Et eksempel på et udfald kunne være de 6 tal: $\{2,3,8,19,21,35\}$. Den rækkefølge, tallene udtrækkes i, er ligegyldig, så $\{2,3,8,19,21,35\}$ er det samme udfald som $\{21,8,35,2,19,3\}$. Når vi skal have netop dette udfald, har vi altså 6 muligheder, når vi trækker den første kugle, fordi vi kan kun bruge et af tallene, 2, 3, 8, 19, 21 og 35. Derefter har vi kun 5 muligheder tilbage, fordi vi allerede har placeret et af de 6 tal i celle nummer 1:

6	5	4	3	2	1

Sandsynligheden for netop dette udfald er således:

$$\frac{\text{antal gunstige udfald}}{\text{antal mulige udfald}} = \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{49 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44} = 7{,}15 \cdot 10^{-8}$$

.

Opgave 1

I et andet Lotto er der 56 kugler i en beholder, og kuglerne er nummereret med tallene $1, 2, \ldots, 56$. Fra beholderen trækkes én kugle ad gangen, og der trækkes i alt syv kugler.

a) Bestem antallet af mulige udfald i dette Lotto.

b) Opskriv et selvvalgt muligt udfald, og bestem sandsynligheden for dette udfald.

Vi indfører nu en stokastisk variabel til at beskrive de forskellige udfald i udfaldsrummet med tal. Populært sagt sætter vi den stokastiske variabel til at "tælle" de *ens* udfald. Hvis vi fx kaster to ens terninger og ser på, at summen af øjentallene giver 7, så er de mulige udfald: $\{1,6\}$, $\{2,5\}$ og $\{3,4\}$.
Da summen af øjentallene skal være 7, så skriver vi $X = 7$. Vi vil senere se, at sandsynligheden for, at $X = 7$ er $\frac{6}{36} = \frac{1}{6}$. Vi skriver $P(X = 7) = \frac{1}{6}$.

Definition 2 **Stokastisk variabel**

I et endeligt sandsynlighedsfelt er en stokastisk variabel X en funktion, der til hvert udfald i U knytter et reelt tal.

Sandsynligheden for, at X antager værdien x_i, skrives som $P(X = x_i)$, og det beregnes som summen af sandsynlighederne for de udfald, der knyttes til x_i.

Til en stokastisk variabel X knytter vi en sandsynlighedstabel, hvor sandsynlighederne for, at den stokastiske variabel antager værdien x_i, betegnes p_i:

Stokastisk variabel $X = x_i$	x_1	x_2	…	…	x_n
Sandsynlighed $P(X = x_i)$	p_1	p_2	…	…	p_n

Bemærk, at betegnelsen p_i er en generel betegnelse for en sandsynlighed, som vi bruger om både et udfalds sandsynlighed og en sandsynlighed for, at en stokastisk variabel antager en bestemt værdi. Vi kan forstå X som en variabel, der kan være lig med n forskellige værdier x_i. Hvis n er 10, så kan den stokastiske variabel X antage 10 forskellige værdier. Hvert af tallene x_i optræder med en sandsynlighed p_i. Hvis vi igen ser på kast med to terninger, hvor X tæller summen af øjentallene, så kan X antage 11 forskellige værdier: $2,3,...,12$. Vi vil senere beregne sandsynligheden knyttet til hver x-værdi.
De enkelte sandsynligheder p_i er tal, der alle ligger i intervallet $[0;1]$, og summen af alle sandsynlighederne p_i er 1. Udtrykket $P(X = 2) = 0,3$ skal læses sådan, at den stokastiske variabel X kan antage værdien 2, dvs. $X = 2$, og sandsynligheden for dette, dvs. $P(X = 2)$, er lig med $0,3$.

Opgave 2 **Kast med to mønter**

Vi kaster nu med to ens mønter og i hvert kast tæller vi antallet af mønter, der viser krone. Her bestemmes eksempelvis sandsynligheden for at få 0 krone, dvs. $P(X = 0)$, ved at tælle de gunstige udfald for dette eksperiment. Da vi kan få 0 krone på én måde, så er antal gunstige udfald lig med 1.

a) Forklar, at X kan være lig med tallene 0,1 og 2.

b) Opskriv de mulige udfald, med betegnelserne $p = plat$ og $k = krone$, dvs. hvis begge mønter viser plat, så betegnes udfaldet pp.

c) Tæl op, og angiv antallet af gunstige udfald for hver af: $X = 0$, $X = 1$ og $X = 2$.

d) Udfyld sandsynlighedstabellen nedenfor, idet du i hvert tilfælde udregner $\frac{\text{antal gunstige udfald}}{\text{antal mulige udfald}}$:

Stokastisk variabel $X = x_i$	0	1	2
Sandsynlighed $P(X = x_i)$			

Middelværdi og spredning

I det følgende ser vi på middelværdi, varians og spredning for en stokastisk variabel. Det er tre tal, som siger noget om den stokastiske variabels sandsynlighedsfordeling. De tre størrelser middelværdi, varians og spredning kender vi allerede fra emnet deskriptiv statistik, og vi kender formler for udregning af de tre størrelser i den sammenhæng. Hvis vi tænker på sandsynlighederne som frekvenser, så svarer middelværdien til et *vægtet gennemsnit* af de tal, som den stokastiske variabel X kan antage. "*Vægtet*" betyder, at man i beregningen tager hensyn til, hvilken sandsynlighed tallene optræder med. Varians og spredning er to værdier, der fortæller noget om, hvor langt de talværdier, som den stokastiske variabel X antager, ligger fra middelværdien.

Definition 3 **Middelværdi af en stokastisk variabel**

Middelværdien af en stokastisk variabel X betegnes μ, og udregnes som det vægtede gennemsnit:

$$\mu = p_1 \cdot x_1 + p_2 \cdot x_2 + ... + p_n \cdot x_n .$$

Man anvender også betegnelsen $E(X)$ for middelværdien af en stokastisk variabel X. Betegnelsen E stammer fra det engelske ord "Expected value", som betyder "forventet værdi". *Store tals lov* siger, at gennemsnittet af rigtig mange udfald af en stokastisk variabel X vil nærme sig middelværdien μ, deraf betegnelsen "den *forventede* værdi".

Eksempel 3 **Kast med en terning – fortsat**

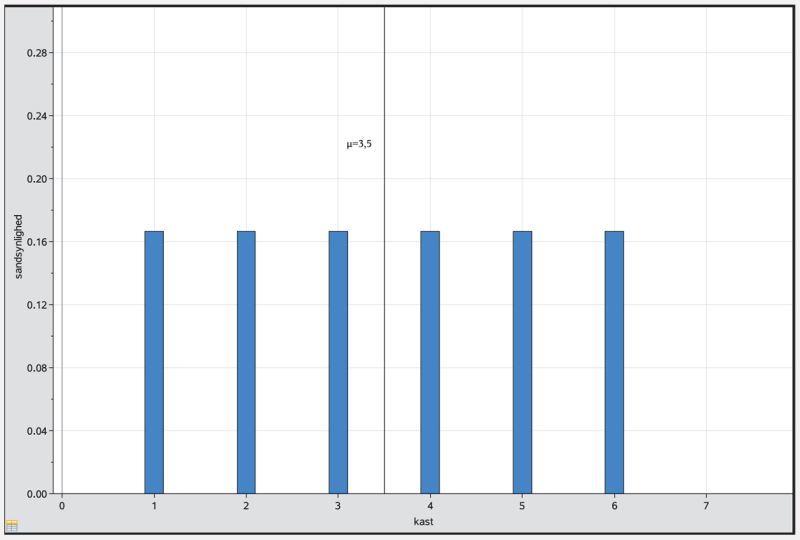
Sandsynlighedstabellen for kast med en terning, hvor den stokastiske variabel X tæller antallet af øjne på terningen, er vist herunder.

Stokastisk variabel $X = x_i$	1	2	3	4	5	6
Sandsynlighed $P(X = x_i)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Middelværdien af X bliver ifølge definitionen:

$$\mu = 1 \cdot \tfrac{1}{6} + 2 \cdot \tfrac{1}{6} + 3 \cdot \tfrac{1}{6} + 4 \cdot \tfrac{1}{6} + 5 \cdot \tfrac{1}{6} + 6 \cdot \tfrac{1}{6} = 3{,}5 .$$

Hvis vi tegner et søjlediagram for X sammen med middelværdien, så får vi:



Eksemplet viser, at $\mu = 3{,}5$ ikke er blandt x_i'erne. Middelværdien kan altså godt antage en værdi, der ikke optræder blandt de mulige værdier for den stokastiske variabel.

Eksempel 4 **Kast med to terninger**

Vi kaster med to terninger og tæller summen af de to terningers øjne. Vi kan illustrere udfaldene i følgende tabel:

Terning 2 / Terning 1	1	2	3	4	5	6
1		$\{1,2\}$				
2	$\{2,1\}$					
3						
4						
5						
6						

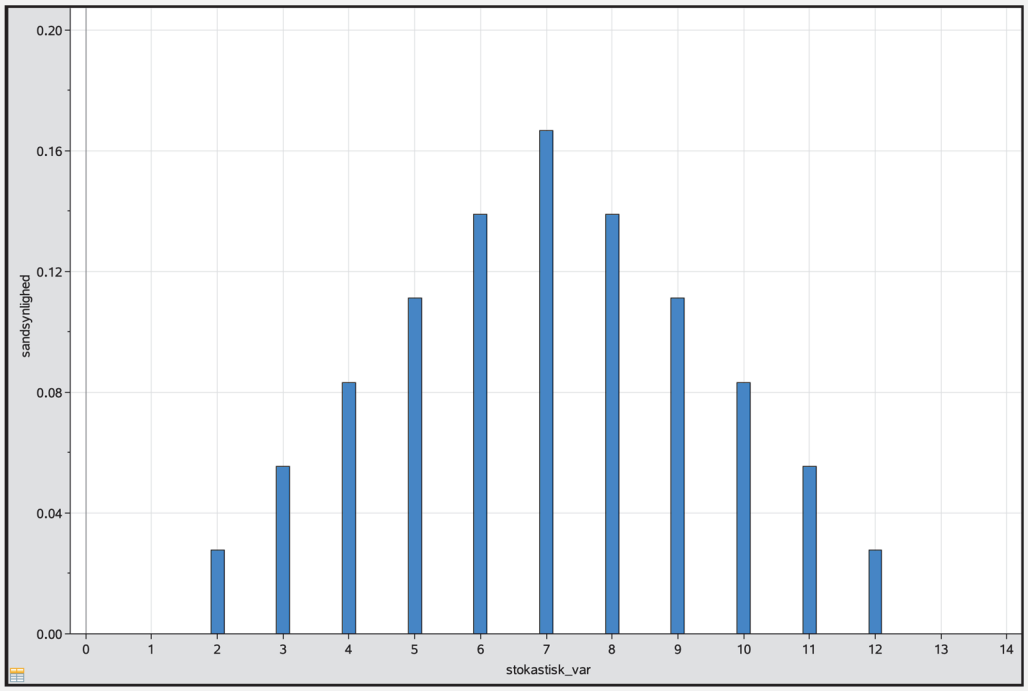
Her bestemmes eksempelvis $P(X = 3)$ ved først at tælle antal gunstige udfald for summen 3. Vi tæller her to, da vi kan få 3 øjne på to måder $\{1,2\}$ og $\{2,1\}$. Vi tæller herefter antal mulige udfald, som her er $6 \cdot 6 = 6^2 = 36$, da der er seks muligheder for hver terning. Hermed bestemmes sandsynligheden for at få summen 3 ved et kast med to terninger:

$$P(X = 3) = \frac{\text{antal gunstige udfald}}{\text{antal mulige udfald}} = \frac{2}{36}.$$

Sandsynlighedsfordelingen for den stokastiske variabel X bliver dermed:

Stokastisk variabel $X = x_i$	2	3	4	5	6	7	8	9	10	11	12
Sandsynlighed $P(X = x_i)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Tegnes et søjlediagram for X får vi.



Opgave 3 Benyt definition 3 til at bestemme middelværdien for denne stokastiske variabel X.

Du skal få $\mu = 7$.

De fleste værktøjsprogrammer har indbyggede kommandoer til at tegne søjlediagram og bestemme middelværdi.

Opgave 4 En stokastisk variabel X, der tæller antallet af krone ved kast med 3 mønter, har følgende sandsynlighedsfordeling.

Stokastisk variabel $X = x_i$	0	1	2	3
Sandsynlighed $P(X = x_i)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

a) Indtast sandsynlighedsfordelingstabellen i dit værktøjsprogram.

b) Undersøg, hvordan man kan tegne søjlediagrammer i dit værktøjsprogram, og tegn et søjlediagram for sandsynlighedsfordelingen i dit værktøjsprogram.

c) Bestem middelværdien for X i dit værktøjsprogram. Du skal få $\mu = 1{,}5$.

Definition 4 **Varians og spredning af en stokastisk variabel**

Variansen, $\mathrm{Var}(X)$, af en stokastisk variabel X med middelværdien μ repræsenterer det gennemsnitlige *afstandskvadrat* til middelværdien μ. Det udregnes som:

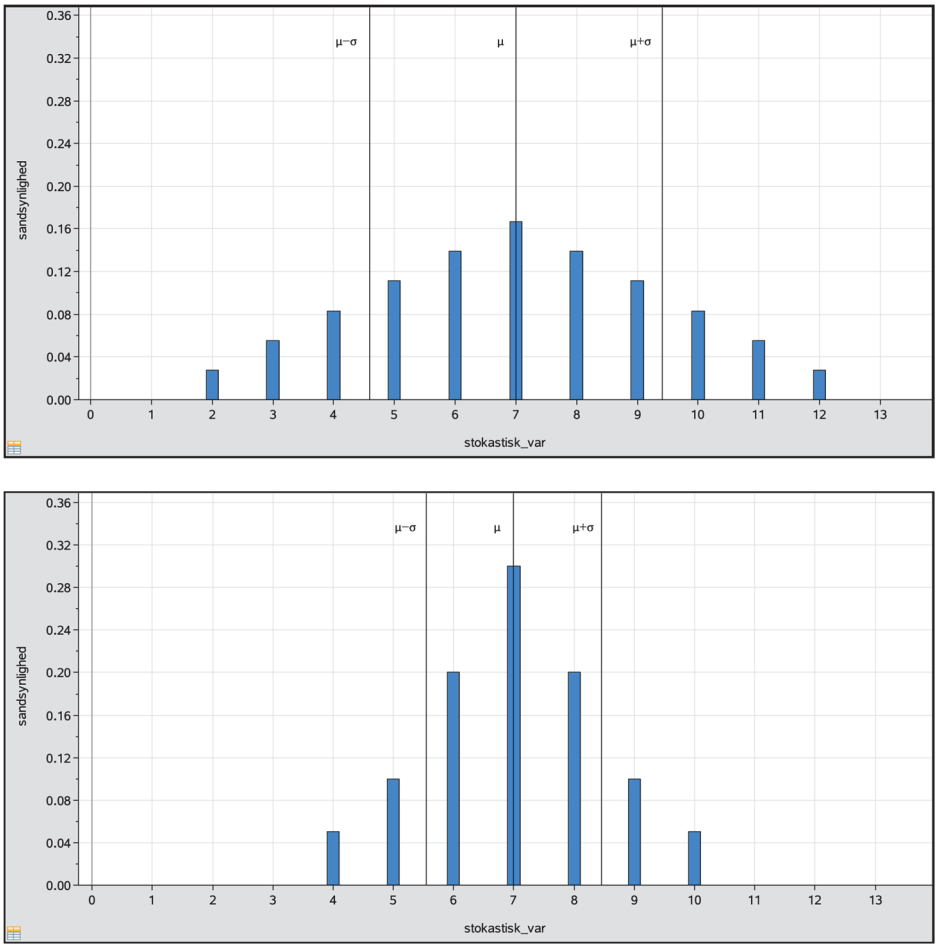
$$\mathrm{Var}(X) = p_1 \cdot (x_1 - \mu)^2 + \ldots + p_n \cdot (x_n - \mu)^2 .$$

Spredningen, σ, udregnes som kvadratroden af variansen:

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

Det gennemsnitlige afstandskvadrat til middelværdien er et mål for, hvor langt den stokastiske variabels talværdier gennemsnitligt ligger fra middelværdien. Værdien $x_i - \mu$ betegner afstanden fra x_i til μ (regnet med fortegn). Vi ønsker ikke negative afstande, så derfor ser vi på afstands*kvadratet*, dvs. vi *kvadrerer* afstanden, så vi får $(x_i - \mu)^2$. Derefter udregner vi det vægtede gennemsnit af disse ved at gange med sandsynlighederne.

Når de stokastiske variables talværdier ligger langt fra middelværdien får vi en høj varians, og omvendt får vi en lille varians, hvis de ligger tæt på. De kvadrerede afstande vægtes med de tilhørende sandsynligheder, hvorved de udfald, der kun optræder med lille sandsynlighed, bidrager mindre til variansen, end de udfald, der optræder med større sandsynlighed. Figuren nedenfor viser to søjlediagrammer for to stokastiske variable med samme middelværdi, men med forskellig varians og dermed forskellig spredning.



Når vi udfører et stokastisk eksperiment, så repræsenterer spredningen i en vis forstand den forventede afstand, et udfald vil have til middelværdien. Dette gælder, fordi vi tager kvadratroden af variansen, når vi skal beregne spredningen. Spredningen er specielt brugbar, når man har to sammenlignelige eksperimenter.

Eksempel 5 **Kast med en terning**

Vi benytter definitionen til at bestemme varians og spredning for den stokastiske variabel X, der tæller antallet af øjne ved kast med én terning.

Vi har tidligere udregnet $\mu = 3{,}5$. Vi udregner variansen for X:

$$\mathrm{Var}(X) = \frac{1}{6}\cdot(1-3{,}5)^2 + \frac{1}{6}\cdot(2-3{,}5)^2 + \frac{1}{6}\cdot(3-3{,}5)^2 + \frac{1}{6}\cdot(4-3{,}5)^2 + \frac{1}{6}\cdot(5-3{,}5)^2 + \frac{1}{6}\cdot(6-3{,}5)^2 = 2{,}92$$

Og herudfra spredningen for X:

$$\sigma = \sqrt{\mathrm{Var}(X)} = \sqrt{2{,}92} = 1{,}71\,.$$

Opgave 5 Benyt definition 4 til at bestemme varians og spredning for den stokastiske variabel, der tæller summen af øjnene ved kast med to terninger.

De fleste værktøjsprogrammer har indbyggede kommandoer til at bestemme varians og spredning.

Opgave 6

a) Undersøg, hvilke kommandoer dit værktøjsprogram anvender til bestemmelse af varians og spredning for en stokastisk variabel X.

b) Benyt programmet til at beregne middelværdi, varians og spredning for en stokastisk variabel X, der angiver summen af øjne ved kast med to terninger.

c) Opstil et regneark i dit værktøjsprogram, der viser sandsynlighedsfordelingen.

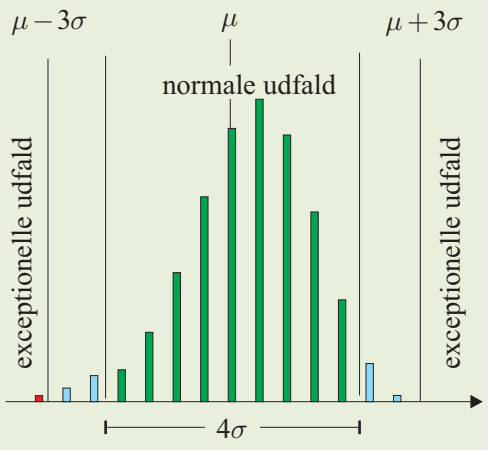
d) Benyt programmet til at tegne et søjlediagram for sandsynlighedsfordelingen.

Når vi udfører et stokastisk eksperiment, så *forventer* vi, at værdierne for den stokastiske variabel ligger inden for en rimelig afstand fra middelværdien. Spredningen angiver, hvor stor denne afstand kan være. Det kan dog ske, at vi får værdier, der ligger langt fra middelværdien, men selvfølgelig ganske sjældent. Vi definerer herunder, hvilke værdier for den stokastiske variabel vi vil anse for *normale*, og hvilke vi vil anse for *exceptionelle*.

Definition 5 **Normale og exceptionelle værdier**

En værdi x for en stokastisk variabel X kaldes *normal*, hvis den ligger inden for to spredninger fra middelværdien μ, dvs. $\mu - 2\sigma \le x \le \mu + 2\sigma$.

En værdi x for en stokastisk variabel X kaldes *exceptionel*, hvis den ligger længere væk end tre spredninger fra middelværdien μ, dvs. $x < \mu - 3\sigma$ eller $x > \mu + 3\sigma$.



Opgave 7 **Kast med to terninger**

Vi kaster med to terninger, og den stokastiske variabel X betegner summen af øjnene.

a) Bestem de værdier for X, der er normale.

b) Bestem de værdier for X, der er exceptionelle.

Binomialfordelingen

I mange praktiske situationer har vi kun to udfald. Disse situationer kan typisk modelleres med binomialfordelingen.

Vi vil i det følgende se nærmere på binomialfordelingen, samt anvendelser heraf.

Eksempler på binomialmodeller kan være et vist antal kast med en ærlig mønt, hvor vi har to udfald; plat og krone. Sandsynligheden for at få plat er $\frac{1}{2}$, hver gang vi kaster mønten. Et andet eksempel er kast med en terning, hvor sandsynligheden for at få en sekser er $\frac{1}{6}$, hver gang vi kaster terningen.

Binomialmodeller er kendetegnet ved, at det samme eksperiment gentages en række gange. Dette eksperiment kaldes for *basiseksperimentet*, og de enkelte basiseksperimenter i modellen antages ikke at påvirke hinanden, vi siger de er *uafhængige*. Et basiseksperiment har netop to mulige udfald, som vi kalder henholdsvis *succes* og *fiasko*. Sandsynligheden for at få succes i et basiseksperiment kaldes *basissandsynligheden*. Basissandsynligheden er således den samme i hvert basiseksperiment.

Vi betragter nu en generel situation, hvor vi laver n uafhængige gentagelser af et basiseksperiment. I hvert forsøg er der to mulige udfald; *succes* og *fiasko*. Sandsynligheden for succes, dvs. basissandsynligheden, som jo er den samme i alle n basiseksperimenter, benævnes p, hvor $0 < p < 1$. Sandsynligheden for fiasko bliver dermed $1 - p$, fordi summen af de to sandsynligheder jo skal være 1.

Vi indfører en stokastiske variabel X, der tæller antallet af succeser. Så er X binomialfordelt, og vi skriver $X \sim b(n, p)$, hvor n kaldes *antalsparameteren*, og p kaldes *sandsynlighedsparameteren*.

Vi repeterer herunder begreberne "fakultet" og kombinationer", som er vigtige elementer i binomialfordelingens sandsynlighedsfunktion.

Definition 6 **Fakultetstallene**

For et naturligt tal n forstås det n'te fakultetstal $n!$ som produktet af de n første hele tal, dvs.

$$n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 3 \cdot 2 \cdot 1.$$

Desuden gælder, at $0! = 1$.

Kombinationer og binomialkoefficienter

For hele tal n og r, hvor $0 \le r \le n$, defineres *binomialkoefficienten* $K(n,r)$ ved

$$K(n,r) = \frac{n!}{r! \cdot (n-r)!}.$$

Eksempel 6 Vi har et sæt med 52 almindelige spillekort.

Hvis vi udtager 13 spillekort blandt de 52 spillekort, så kan vi udregne antallet af forskellige måder, som vi kan få de 13 spillekort på, ved

$$K(52,13) = \frac{52!}{13! \cdot (52-13)!} = \frac{52!}{13! \cdot (39)!} = 635013559600$$

En af disse mange delmængder kan være delmængden med disse 13 spillekort, som vi har taget blandt alle 52 spillekort:

Sætning 1 $K(n,r)$ angiver hvor mange delmængder med r elementer, der kan udtages af en mængde med n elementer.

Bevis

Vi beviser dette ved at sætte r elementer fra en mængde på n elementer i rækkefølge på to forskellige måder.

Metode 1:

Vi antager først, at vi skal vælge blandt n elementer og sætte disse i rækkefølge på r pladser. På første plads kan man vælge mellem alle n elementer, på anden plads mellem $n-1$ elementer, på tredje plads mellem $n-2$ elementer, osv. Efter den næstsidste plads har vi brugt $r-1$ elementer, og der er derfor $n-(r-1)$ elementer at vælge mellem til den sidste plads.

Plads	1	2	3	…	r
Antal elementer	n	$n-1$	$n-2$	…	$n-(r-1)$

Der er altså i alt $n\cdot(n-1)\cdot(n-2)\cdot\ldots\cdot(n-(r-1))$ forskellige måder at sætte r elementer fra en mængde på n elementer i rækkefølge på. Vi kan omskrive dette udtryk til:

$$n\cdot(n-1)\cdot(n-2)\cdot\ldots\cdot(n-(r-1))=$$

$$n\cdot(n-1)\cdot(n-2)\cdot\ldots\cdot(n-(r-1))\cdot\frac{(n-r)!}{(n-r)!}=$$

Ganger udtrykket $\frac{(n-r)!}{(n-r)!}$, som jo er 1, på

$$\frac{n\cdot(n-1)\cdot(n-2)\cdot\ldots\cdot(n-(r-1))\cdot(n-r)!}{(n-r)!}=$$

Ganger op i tælleren

$$\frac{n\cdot(n-1)\cdot(n-2)\cdot\ldots\cdot(n-(r-1))\cdot(n-r)\cdot(n-(r+1))\cdot\ldots\cdot3\cdot2\cdot1!}{(n-r)!}=$$

Skriver $(n-r)!$ ud

$$\frac{n!}{(n-r)!}$$

Udnytter at
$n!=n\cdot(n-1)\cdot\ldots\cdot(n-(r-1))\cdot(n-r)\cdot(n-r-1)\cdot\ldots\cdot2\cdot1$

Antallet af forskellige måder vi kan sætte r elementer fra en mængde på n elementer i rækkefølge på kan altså beregnes ved $\frac{n!}{(n-r)!}$.

Metode 2:

Vi antager, at vi først vælger *r* elementer blandt de *n* elementer, og herefter sætter dem i rækkefølge på de *r* pladser. Når de *r* elementer, vi har valgt, skal sættes i rækkefølge, kan man på den første plads vælge mellem alle de *r* elementer, på anden plads mellem $r-1$ osv. På den sidste plads er der kun 1 element tilbage at vælge imellem. Der er altså i alt $r\cdot(r-1)\cdot...\cdot 1=r!$ måder at sætte de *r* elementer i rækkefølge på.

Men de *r* elementer kan jo vælges på flere måder, og vi vil antage, at *r* elementer kan udvælges blandt *n* elementer på *x* måder. Det betyder så, at der vil være i alt $x\cdot r!$ måder at sætte *r* elementer fra en mængde på *n* elementer i rækkefølge på.

Sammenligning af metode 1 og 2:

I hver af de to metoder har vi opskrevet to udtryk for det samme. Disse må derfor give samme resultat, og vi får derfor, at:

$x\cdot r!=\frac{n!}{(n-r)!}$	Sætter de to udtryk lig med hinanden
$x=\frac{n!}{r!\cdot(n-r)!}$	Dividerer med $r!$ på begge sider
$x=K(n,r)$	Udnytter definition 6

Da *x* er antallet af måder, hvorpå man kan udvælge *r* elementer blandt *n* elementer, er det ønskede bevist, dvs. at man med binomialkoefficienten $K(n,r)$ kan beregne antallet af måder hvorpå, vi kan udvælge en delmængde med *r* elementer fra en mængde med *n* elementer.

Opgave 8 Binomialkoefficienter kan nemt beregnes med de fleste værktøjsprogrammer.

a) Undersøg, hvordan dit værktøjsprogram beregner binomialkoefficienter, og beregn $K(20,5)$, $K(20,0)$ og $K(20,19)$.

I en binomialmodel kan sandsynligheden for, at der optræder *r* succeser ud af i alt *n* mulige, beregnes ved hjælp af binomialkoefficienter:

Sætning 2 Hvis $X\sim b(n,p)$, så beregnes binomialsandsynligheder således:

$$P(X=r)=K(n,r)\cdot p^r\cdot(1-p)^{n-r},$$

hvor *n* er antalsparameteren, og *p* er sandsynlighedsparameteren.

Bevis

$P(X=r)$ angiver sandsynligheden for, at der optræder *r* succeser ud af de *n* mulige i *n* basiseksperimenter. De basiseksperimenter, der ikke resultater i succes, resulterer jo i fiasko, så derfor må der optræde $n-r$ fiaskoer. Sandsynligheden for succes i basiseksperimentet er *p*, og sandsynligheden for fiasko er så $1-p$.

En måde, hvorpå vi kan få netop *r* succes'er og dermed netop $n-r$ fiasko'er, er ved at vi får succes de første *r* gange og fiasko de sidste $n-r$ gange i de *n* basiseksperimenter. Ser vi på netop den situation, så, så kan vi udregne sandsynlighederne således:

Sandsynligheden for succes de første *r* gange er: $\underbrace{p\cdot p\cdot...\cdot p}_{r\,\text{gange}}=p^r$

Sandsynligheden for fiasko de næste $n-r$ gange er: $\underbrace{(1-p)\cdot(1-p)\cdot...\cdot(1-p)}_{n-r\,\text{gange}}=(1-p)^{n-r}$

Derfor er sandsynligheden for udfaldet: "*r* succes'er og $n-r$ fiasko'er" samlet set $p^r\cdot(1-p)^{n-r}$.

Uanset, på hvilken måde vi trækker de r elementer, så er sandsynligheden givet ved det samme udtryk: $p^r \cdot (1-p)^{n-r}$ (overvej!). Vi vil derfor kunne bestemme den samlede sandsynlighed for at opnå netop *r* succes'er ved at lægge sandsynlighederne for hver af måderne sammen.

Men vi ved fra sætning 1, at antallet af forskellige måder, hvorpå man kan trække *r* elementer ud af en mængde med *n* elementer, kan beregnes med $K(n,r)$. Dvs. når vi tager hensyn til antallet af måder, hvorpå vi kan trække de *r* elementer, så bliver den samlede sandsynlighed for at opnå *r* succeser:

$$P(X=r) = K(n,r) \cdot p^r \cdot (1-p)^{n-r} \ .$$

Hermed er det ønskede bevist.

Eksempel 7 Der trækkes et kort blandt 52 almindelige spillekort. Det noteres, om kortet er et billedkort, hvorefter kortet lægges tilbage i bunken. Dette kan betragtes som basiseksperimentet i en binomialfordeling, hvor basissandsynligheden er $p = \frac{12}{52} = \frac{3}{13}$.

Basiseksperimentet gentages 6 gange, og vi indfører den stokastiske variabel *X*, der tæller antallet af billedkort.

Sandsynlighedsfunktionen for binomialmodellen bliver derfor:

$$P(X=r) = K(6,r) \cdot \left(\frac{3}{13}\right)^r \cdot \left(1-\frac{3}{13}\right)^{6-r} \ .$$

Opgave 9 Udfyld sandsynlighedstabellen for den stokastiske variabel *X*, der tæller antallet af billedkort i binomialmodellen $b(6,\frac{3}{13})$ fra ovenstående eksempel.

$X=r$	0	1	2	3	4	5	6
$P(X=r)$							

Eksempel 8 **Kast med en mønt 20 gange**

Hvis vi kaster en ærlig mønt 20 gange, og lader den stokastiske variabel *X* tælle antallet af "krone", så er *X* binomialfordelt med antalsparameter $n=20$ og sandsynlighedsparameter $p=0,5$.

Vi kan nu bestemme sandsynligheder for forskellige udfald ved hjælp af binomialfordelingens sandsynlighedsfunktion.

Sandsynligheden for, at ingen kast giver "krone":

$$P(X=0) = K(20,0) \cdot 0,5^0 \cdot 0,5^{20} = \frac{20!}{0!20!} \cdot 1 \cdot 0,5^{20} = 0,5^{20} = 9,537 \cdot 10^{-7}.$$

Dvs. sandsynligheden for, at ingen af de 20 kast giver "krone" – eller med andre ord, sandsynligheden for, at *alle* de 20 kast giver "plat", er 0,00009537%.

Sandsynligheden for, at ét kast giver "krone":

$$P(X=1)=K(20,1)\cdot 0{,}5^{1}\cdot 0{,}5^{20-1}=\frac{20!}{1!19!}\cdot 0{,}5\cdot 0{,}5^{19}=20\cdot 0{,}5^{20}=1{,}907\cdot 10^{-5}.$$

Dvs. sandsynligheden for, at ét kast giver "krone" – eller med andre ord, sandsynligheden for, at 19 kast giver "plat" – er 0,001907%.

I de matematiske værktøjsprogrammer er binomialsandsynligheder lagt ind som en kommando. Typisk betegnes kommandoen *BinomPdf* eller *binpdf*.

Opgave 10

a) Benyt dit værktøjsprogram til at beregne følgende sandsynligheder, når X er binomialfordelt med $n=20$ og $p=0{,}5$:

$$P(X=2),\ P(X=10) \text{ og } P(X=11).$$

Eksempel 9 **Kast med en mønt 20 gange (fortsat)**

Sandsynligheden for at få højst 3 "krone" kan naturligvis beregnes som summen af sandsynlighederne for at få hhv. 0, 1, 2 og 3 "krone":

$$P(X\le 3)=P(X=0)+P(X=1)+P(X=2)+P(X=3).$$

Tilsvarende kan man beregne sandsynligheden for at få flere end 15 "krone" ved at summere sandsynlighederne for at få hhv. 16, 17, 18, 19 og 20 "krone":

$$P(X>15)=P(X=16)+P(X=17)+P(X=18)+P(X=19)+P(X=20).$$

Eller alternativt som 1 minus sandsynligheden for at få højst 15 "krone":

$$P(X>15)=1-P(X\le 15).$$

Opgave 11 Udregn sandsynlighederne i ovenstående eksempel i dit matematiske værktøjsprogram.

Opgave 12 En stokastisk variabel X er binomialfordelt med antalsparameter $n=45$ og sandsynlighedsparameter $p=0{,}3$.

a) Bestem $P(X=10)$.

b) Bestem $P(X<5)$.

c) Bestem sandsynligheden for at få færre end 10 succeser.

d) Bestem sandsynligheden for at få flere end 15 succeser.

Opgave 13 Vi kaster en ærlig terning 60 gange, og lader den stokastiske variabel X tælle det antal gange, hvor vi får en sekser.

a) Hvorfor kan dette eksperiment beskrives ved en binomialmodel?

b) Hvad er antalsparameteren og sandsynlighedsparameteren her?

c) Bestem sandsynligheden for at få flere end 20 seksere.

d) Opstil en sandsynlighedsfordelingstabel for X i et regneark.

e) Tegn et søjlediagram for sandsynlighedsfunktionen.

f) Hvilken af søjlerne er højest?

Opgave 14 En undersøgelse viser, at 4% af alle danskere er vegetarer. Blandt danskere udtages en stikprøve på 1035 personer. Da stikprøven er meget lille i forhold til populationen, kan dette behandles ved hjælp af en binomialmodel (overvej dette!).

a) Indfør en passende stokastisk variabel X, og opstil en binomialmodel for antallet af vegetarer i stikprøven.

b) Bestem sandsynligheden for, at der højst er 25 vegetarer i stikprøven.

Opgave 15 X er en binomialfordelt stokastisk variabel. Udfyld i et regneark i et matematisk værktøjsprogram en sandsynlighedsfordelingstabel for X, når $p = 0{,}1$ og $n = 100$.

a) Bestem ved hjælp af værktøjsprogrammet middelværdien for X ud fra sandsynlighedstabellen.

Andre binomialfordelte stokastiske variable X har sandsynlighedsparameter $p = 0{,}1$ og antalsparameter som vist i nedenstående tabel.

n	10	20	40	80	100
p	0,1	0,1	0,1	0,1	0,1
Middelværdi					

b) Bestem middelværdien på samme som før, og udfyld en tabel som ovenfor.

c) Formulér en formodning om bestemmelse af middelværdien for en binomialfordelt stokastisk variabel X ud fra mønsteret i din tabel.

d) Andre binomialfordelte stokastiske variable X har antalsparameter $n = 10$ og sandsynlighedsparameter som vist i nedenstående tabel.

n	10	10	10	10	10
p	0,2	0,3	0,4	0,5	0,6
Middelværdi					

e) Bestem på samme måde som i b) middelværdien og udfyld en ny tabel som ovenfor.

f) Gælder formodningen om bestemmelse af middelværdien for en binomialfordelt stokastisk variabel X ud fra tilfældene i din nye tabel fra d)?

For binomialfordelinger gælder følgende sætning om middelværdi.

Sætning 3 **Middelværdi om binomialfordelingen**

$$\mu = p \cdot n,$$

hvor n er antalsparameteren, og p er sandsynlighedsparameteren.

Beviset for sætningen er mest for A-niveau og kan findes i appendix.

Eksempel 10 **Stikprøve i mobiltelefonproduktion**

Erfaringen har vist, at 3% af mobiltelefonerne i en produktion er defekte. I det følgende antages det, at sandsynligheden for, at en tilfældig mobiltelefon er defekt, er 0,03. Der udtages en stikprøve på 950 mobiltelefoner fra produktionen.

Vi lader X tælle antallet af mobiltelefoner i stikprøven, der er defekte.

Det gennemsnitlige antal defekte mobiltelefoner i stikprøver af denne størrelse kan vi udregne ved

$$\mu = 950 \cdot 0{,}03 = 28{,}5.$$

Vi vil derfor forvente, at der er omkring 28 eller 29 mobiltelefoner, der er defekte.

Sætning 4 **Varians og spredning for binomialfordelingen**

Hvis X er binomialfordelt med antalsparameter n og sandsynlighedsparameter p, så er varians og spredning for X givet ved**:**

$$V(X) = n \cdot p \cdot (1-p)$$

$$\sigma(X) = \sqrt{V(X)} = \sqrt{n \cdot p \cdot (1-p)}.$$

Beviset for sætningen om variansen og spredningen for en binomialfordeling overspringes her.

Opgave 16 En ærlig mønt kastes 3 gange. Den stokastiske variabel X tæller antallet af "krone".

a) Gør rede for, at sandsynlighedstabellen ser ud som vist i tabellen:

Antal krone $X = x_i$	0	1	2	3
Sandsynlighed $P(X = x_i)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

b) Bestem middelværdi og spredning for antal "krone".

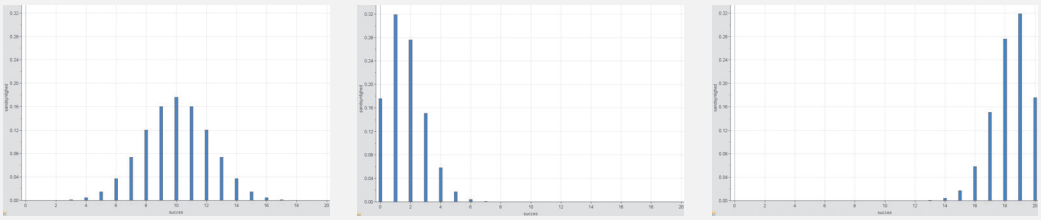
c) Hvilke antal "krone" er normale udfald?

d) Gør rede for, at der ikke findes exceptionelle udfald i dette eksperiment.

Eksempel 11 **Typer af søjlediagrammer for binomialfordelinger**

På figuren herunder ses søjlediagrammer for binomialfordelingens sandsynlighedsfunktion for henholdsvis $b(20,\frac{1}{2})$, $b(20,\frac{1}{12})$ og $b(20,\frac{11}{12})$.

Søjlediagrammerne er tegnet på baggrund af regneark med sandsynligheder for de enkelte udfald som i opgave 13.



$b(20,\frac{1}{2})$ $b(20,\frac{1}{12})$ $b(20,\frac{11}{12})$

Vi ser, at de grafiske billeder er ret forskellige. Hvis p er lille, får søjlediagrammet en hale til højre, og fordelingen kaldes *højreskæv*. Omvendt, hvis p er stor, får vi en hale til venstre, og fordelingen kaldes *venstreskæv*. I alle andre tilfælde har vi en type som den venstre illustration. Disse typer kaldes for *centrale* fordelinger.

Vi så ovenfor hvordan henholdsvis centrale og skæve fordelinger kunne identificeres ud fra søjlediagram for deres sandsynlighedsfunktioner. Vi kan også definere skævhed med udgangspunkt i sandsynlighedsfordelingens middelværdi. Bemærk, at skævhed kun er interessant for store værdier af n.

Definition 7 **Højreskæv og venstreskæv fordeling**

En binomialfordeling kaldes *højreskæv*, hvis middelværdien μ er mindre end 5.

En binomialfordeling kaldes *venstreskæv*, hvis middelværdien μ er større end $n-5$.

Øvrige binomialfordelinger kaldes *centrale*.

Opgave 17

a) Opstil et regneark i dit matematiske værktøjsprogram for en binomialfordelt stokastisk variabel X, hvor antalsparameteren $n=20$ og sandsynlighedsparameteren p bestemmes med en skyder.

b) Tegn et søjlediagram for sandsynlighedsfunktionen.

c) Varierer p. For hvilke p er fordelingen venstreskæv?

d) Varierer p. For hvilke p er fordelingen højreskæv?

e) Varierer p. For hvilke p er fordelingen central?

f) Er der overensstemmelse mellem definition 7 og eksempel 6?

Som vi tydeligt så ovenfor, så ser søjlediagrammerne for binomialfordelingens sandsynlighedsfunktion meget forskellige ud for forskellige værdier af p. Nogle er symmetriske omkring middelværdien, mens andre er skæve.

Uanset om binomialfordelingen er symmetrisk eller skæv, gælder der altid, at sandsynligheden topper lige omkring middelværdien.

Sætning 5 **Mest sandsynlige udfald for en binomialfordelt stokastisk variabel**

Hvis middelværdien er et helt tal, er middelværdien det mest sandsynlige udfald.

Hvis middelværdien ikke er et helt tal, er det mest sandsynlige udfald én af de to heltalsnaboer til middelværdien.

Denne sætning bevises ikke her.

Eksempel 12 **Stikprøve af mobiltelefoner (fortsat)**

Vi lader X være antallet af mobiltelefoner i stikprøven, der er defekte. Antalsparameteren er $n = 950$ og sandsynlighedsparameteren er $p = 0,03$.

Ovenfor fandt vi middelværdien $\mu = 28,5$ som ikke er et helt tal.

Vi udregner sandsynlighederne

$$P(X = 28) = K(950,28) \cdot 0,03^{28} \cdot (1-0,03)^{950-28} = 0,076.$$

$$P(X = 29) = K(950,29) \cdot 0,03^{29} \cdot (1-0,03)^{950-29} = 0,075.$$

Det mest sandsynlige udfald er dermed 28 defekte mobiltelefoner.

Opgave 18 Et gartneri oplyser, at for en bestemt slags forårsløg vil 9 ud af 10 løg spire. En husejer køber 45 af de pågældende forårsløg.

a) Bestem middelværdi og spredning for antallet af spirende forårsløg.

b) Bestem det mest sandsynlige antal spirende forårsløg.

c) Bestem sandsynligheden for, at mindst 40 af forårsløgene vil spire.

Opgave 19 Sandsynligheden for, at en LED pære brænder mere end 25000 timer, er 87%. En husejer skifter 40 lampepærer til LED pærer.

a) Indfør en binomialfordelt stokastisk variabel X, der betegner antallet af LED pærer, der brænder mere 25000 timer.

b) Bestem middelværdi for antallet af LED pærer, der brænder mere end 25000 timer.

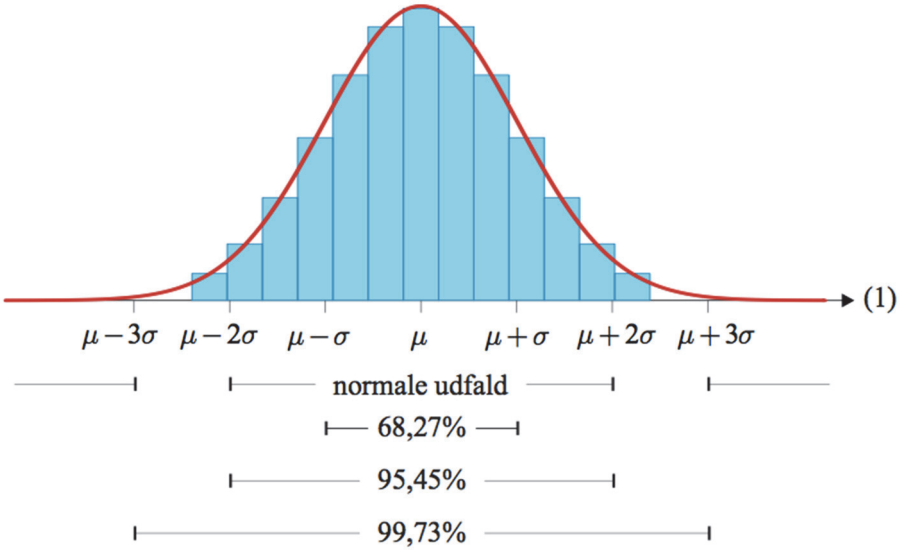
c) Bestem det mest sandsynlige antal LED pærer, der brænder mere end 25000 timer.

Andre sandsynlighedsfordelinger

Normalfordelingen

Normalfordelingen er den mest udbredte statistiske fordeling, blandt andet fordi den repræsenterer en god tilnærmelse til mere komplicerede fordelinger som fx binomialfordelingen.

Normalfordelingen er kontinuert i modsætning til binomialfordelingen som er diskret. Den røde graf herunder viser normalfordelingens karakteristiske klokkeform. Vi vil senere få brug for at vide, at normalfordelte stokastiske variable er fordelt som vist nedenfor.



Som vist på figuren fordeler observationer knyttet til normalfordelte stokastiske variable sig sådan, at:

- ca. 68% af observationerne forventes at ligge inden for én spredning fra middelværdien.
- ca. 95% af observationerne forventes at ligge inden for to spredninger fra middelværdien (dvs. de er normale).
- ca. 0,25% af observationerne forventes at ligge uden for tre spredninger fra middelværdien (dvs. de er exceptionelle).

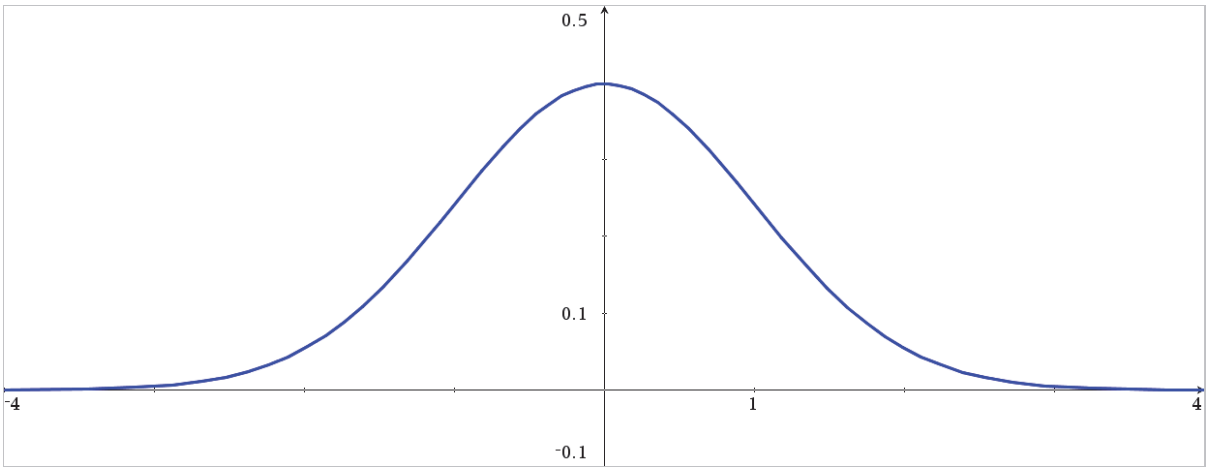
Når en stokastisk variabel Z er normalfordelt med middelværdi μ og spredning σ skriver vi: $Z \sim N(\mu,\sigma)$.

Den mest simple af alle normalfordelinger kaldes *standardnormalfordelingen*, og er defineret ved at have middelværdi $\mu = 0$ og spredning $\sigma = 1$, dvs. $Z \sim N(0,1)$.

Standardnormalfordelingen har en sandsynlighedsfunktion med den ret komplicerede forskrift

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\cdot x^2}.$$

Grafen for standardnormalfordelingens sandsynlighedsfunktion har tilsvarende den karakteristiske klokkeform:



Værktøjsprogrammerne har indbyggede kommandoer for sandsynlighedsfunktionen, typisk betegnes disse *NormPdf*.

Vi kan benytte normalfordelingen som tilnærmelse til binomialfordelingen med samme middelværdi og spredning, hvis vi har en *central* binomialfordeling, dvs. hvis $5 < \mu < n - 5$.

Sandsynlighedsfunktionen for den normalfordelte stokastiske variabel Z, der i disse tilfælde tilnærmer den tilsvarende binomialfordelte stokastiske variabel X, med middelværdi μ og spredning σ har forskriften

$$\varphi(z) = \frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot \mathrm{e}^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2}.$$

Eksempel 13 En binomialfordelt stokastiske variabel X har sandsynlighedsparameteren $p = 0{,}4$ og antalsparameteren $n = 20$, og dermed middelværdi:

$$\mu = 0{,}4\cdot 20 = 8$$

og spredning

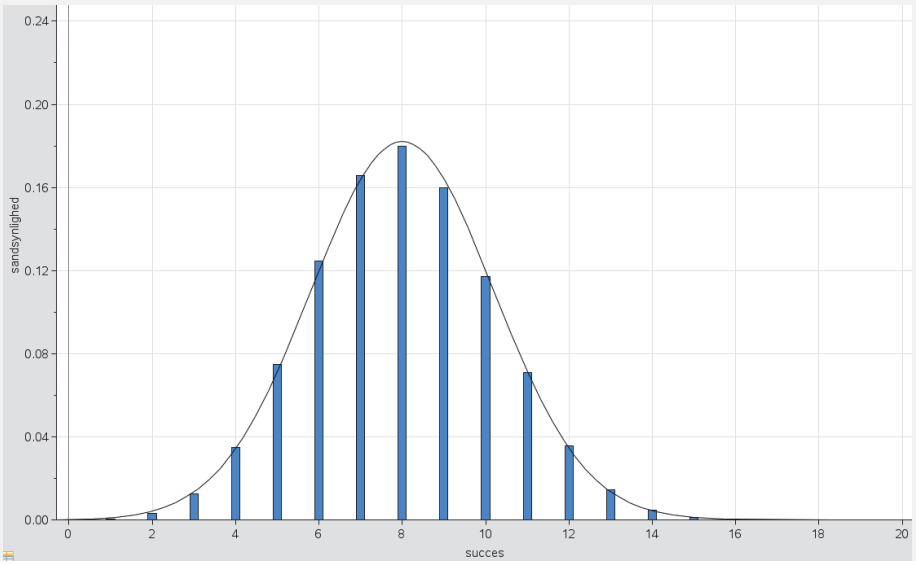
$$\sigma = \sqrt{20\cdot 0{,}4\cdot(1-0{,}4)} = 2{,}1909.$$

Sandsynlighedsfunktionen for den normalfordelte stokastiske variabel Z, der tilnærmer den binomialfordelte stokastiske variabel X har dermed forskriften

$$\varphi(z) = \frac{1}{\sqrt{2\pi}\cdot 2{,}1909}\cdot \mathrm{e}^{-\frac{1}{2}\left(\frac{z-8}{2{,}1909}\right)^2}.$$

Tilnærmelsen kan benyttes, da middelværdien $\mu = 8$ ligger mellem 5 og $20-5=15$.

Graferne for de to sandsynlighedsfunktioner er her tegnet i samme koordinatsystem:



Opgave 20

a) Tegn et søjlediagram for sandsynlighedsfunktionen for en binomialfordelt stokastisk variabel X i dit matematiske værktøjsprogram for tre forskellige selvvalgte binomialfordelinger $b(n,p)$, idet du først opstille tabeller for disse.

b) Tegn et søjlediagram for sandsynlighedsfunktionen for en normalfordelt stokastisk variabel Z i samme koordinatsystem, hvor middelværdien er $\mu = n\cdot p$ og spredningen er $\sigma = \sqrt{n\cdot p\cdot(1-p)}$, dvs. middelværdien og spredningen for Z bestemmes ud fra de samme værdier for n og p som ovenfor.

c) Undersøg, ved at vælge middelværdier under 5 og større end $n-5$, hvorfor det kræves, at middelværdien skal ligge mellem 5 og $n-5$.

Opgave 21

a) Vælg en værdi for n og p, så middelværdien μ ligger mellem 5 og $n-5$.

b) Tegn i dit matematiske værktøjsprogram en lodret linje l med ligningen $x = n \cdot p - 1{,}96 \cdot \sqrt{n \cdot p \cdot (1-p)}$.

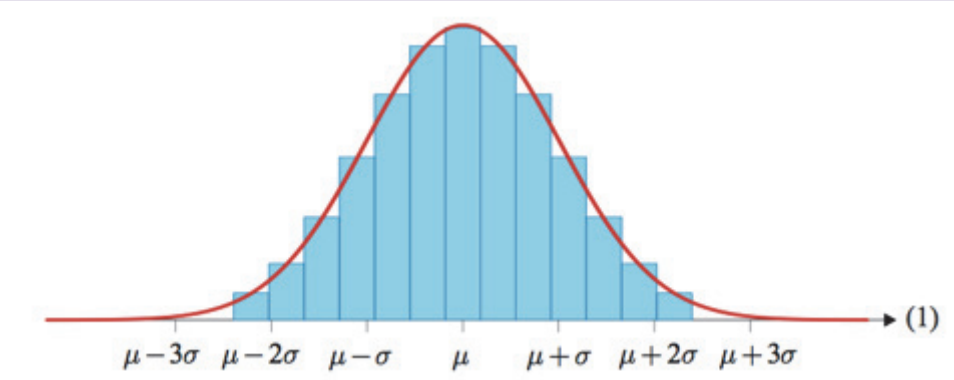
c) Tegn i dit matematiske værktøjsprogram en lodret linje m med ligningen $x = n \cdot p + 1{,}96 \cdot \sqrt{n \cdot p \cdot (1-p)}$.

Grafen for φ afgrænser sammen med linjerne l, m og førsteaksen en punktmængde M.

d) Benyt dit matematiske værktøjsprograms indbyggede grafiske kommando til at bestemme arealet af M.

e) Hvilken værdi har arealet af M?

f) Sammenlign din figur og værdi af arealet med figuren fra tidligere, hvor du fokuserer $\mu + 2 \cdot \sigma$ og $\mu - 2 \cdot \sigma$.



Normalfordelingen vil ikke blive selvstændigt behandlet yderligere i dette materiale.

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 "plat", 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: *Mønten er ærlig, dvs.* $p = \frac{1}{2}$.

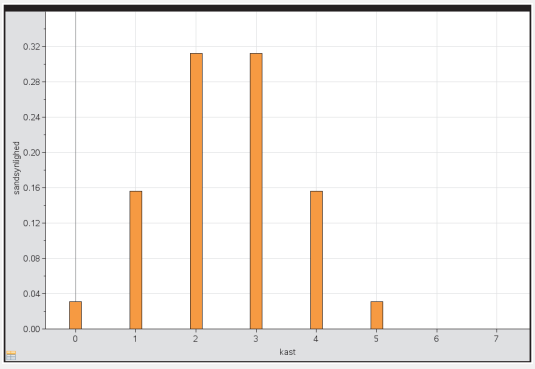
H_1: *Mønten er uærlig, dvs.* $p \neq \frac{1}{2}$.

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:

$$H_0\colon p=p_0$$

Acceptområdet betegnes $\{k_v,\ldots,k_h\}$, 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 $P(X\le k_v)>0{,}025$.

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

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.* $p\le\frac{1}{2}$.

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 $\{0,...,k_h\}$, 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:

Værdien k_h bestemmes som den største værdi af X, 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 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: *Terningen giver toere en fjerdedel af gangene, dvs.* $p = \frac{1}{4}$.

H_1: *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 hypotesetestet være ensidet, nemlig højresidet, svarende til, at højst en tredjedel af forsøgspersonerne vil vælge "den rigtige".

Nulhypotesen 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 forsøgsperosnerne kan smage forskel i hver af smagningerne. Lader vi p betegne sandsynligheden for, at en forsøgsperson kan smage forskel, er vores nulhypotese og alternative hypotese:

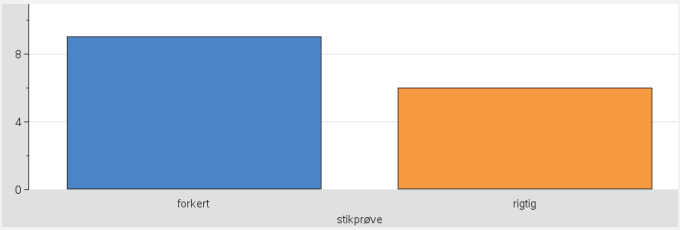
$$H_0\colon\ p \le \frac{1}{3}. \qquad\qquad H_1\colon\ p > \frac{1}{3}.$$

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.

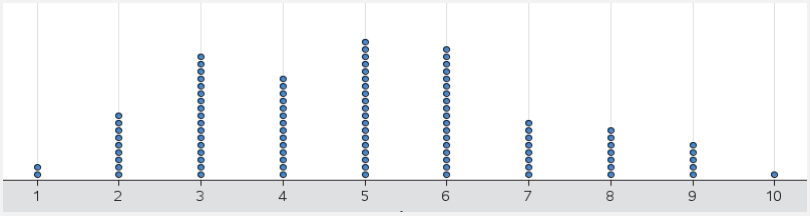
For at undersøge, om det er et udtryk for tilfældig variation, simulerer vi nu situationen i stedet for at foretage teoretiske beregninger i binomialfordelingen.

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.

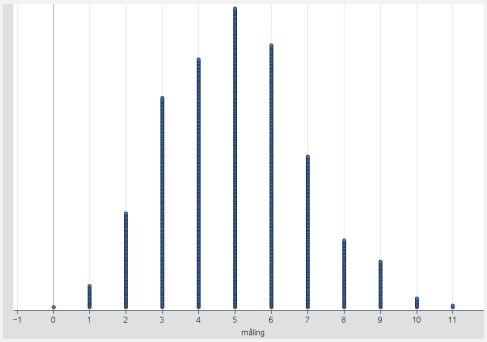


Vi lader nu værktøjsprogrammet udtage en sådan stikprøve med 15 forsøgspersoner rigtig mange gange, og vi noterer hver gang, hvor mange af forsøgspersonerne, der vælger "den rigtige" vingummi.

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 $X=8$.

a) Opstil en tabel over sandsynlighederne.

b) Tegn et søjlediagram for sandsynlighedsfunktionen.

c) Beregn $P(X=8)$.

d) Beregn p-værdien.

e) Bestem den kritiske mængde.

f) 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.

a) Opstil en nulhypotese, når det antages at testet er dobbeltsidet.

b) 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.

a) Formuler en nulhypotese, hvor udgangspunktet er, at partiets vælgertilslutning er uændret.

b) Afgør om testet af nulhypotesen skal være tosidet eller ensidet.

c) Opbyg en simulering af nulhypotesen i et matematisk værktøjsprogram.

d) Bestem teststørrrelsen, dvs. den observerede værdi.

e) Viser simuleringen den samme vælgertilslutning, som Gallups undersøgelse vist? Eller er den større?

f) Udfør nu 1000 simuleringer af nulhypotesen.

g) Bestem antallet af simuleringer, der viser den samme eller en større værdi for vælgertilslutningen.

h) Bestem den eksperimentelle p-værdi, og sammenlign med signifikansniveauet.

Konfidensintervaller

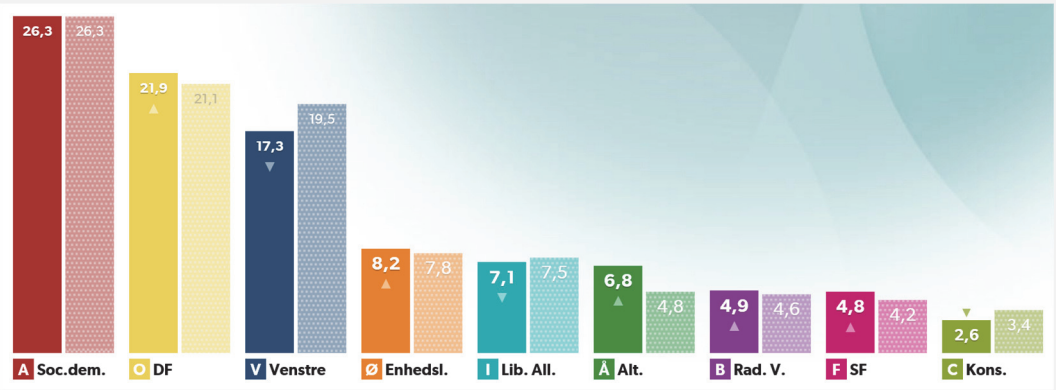
Der offentliggøres jævnligt analyser, der viser, hvilke politiske partier der går frem eller tilbage.

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, ligge 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

a) 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.

b) Opret en stikprøve på 200 værdier af den stokastiske variabel X i regnearket.

c) Bestem middelværdien og spredningen for de 200 værdier.

d) Bestem konfidensintervallet for den sande værdi af p ud fra middelværdi og spredning.

e) Afgør om værdien 0,21 ligger i konfidensintervallet.

f) Tegn konfidensintervallet som et linjestykke i dit værktøjsprogram.

g) Gennemfør trinene a) - f) 100 gange.

h) Bestem andelen af konfidensintervaller, der indeholder værdien 0,21.

i) 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.

a) Tegn grafen for $+u$ som funktion af $\hat{p}$ i et passende interval.

b) Ved hvilken vælgertilslutning er usikkerheden størst?

c) For hvilket parti er usikkerheden størst?

d) 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%.

a) Bestem ud fra stikprøven i 2016 et 95% konfidensinterval for andelen af voksne danskere, der regelmæssigt dyrker motion.

b) 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.

Appendix

1. Hypergeometrisk fordeling

En population på n elementer består af netop to typer elementer: succes og fiasko.

Vi udtager heraf en stikprøve på k elementer af populationen. Lad os antage, at der er a succeser i populationen. Antallet af fiaskoer er dermed $n-a$.

Vi siger, at den stokastiske variabel X der tæller antallet af succeser i stikprøven er *hypergeometrisk* fordelt, og vi skriver $X \sim h(k,a,n)$.

Sætning 7 **Den hypergeometriske sandsynlighedsfunktion**

$$P(X=x)=\frac{\binom{a}{x}\cdot\binom{n-a}{k-x}}{\binom{n}{k}}, \qquad 0\le x\le a \text{ og } 0\le x\le k.$$

Her betegner n det samlede antal elementer, a betegner antal succeser, og k betegner det antal elementer, der skal udtages blandt de n elementer.

Bemærk, at sandsynlighedsfunktionen også kan skrives som

$$P(X=x)=\frac{K(a,x)\cdot K(n-a,k-x)}{K(n,k)}.$$

For den hypergeometriske fordeling gælder følgende sætninger om middelværdi, varians og spredning. Sætningerne bevises ikke her.

Sætning 8 **Middelværdi for den hypergeometriske fordeling**

$$E(X)=\frac{k\cdot a}{n}$$

Sætning 9 **Varians og spredning for den hypergeometriske fordeling**

$$V(X)=\frac{k\cdot a\cdot(n-k)\cdot(n-a)}{(n-1)\cdot n^2}$$

$$\sigma(X)=\sqrt{V(X)}=\sqrt{\frac{k\cdot a\cdot(n-k)\cdot(n-a)}{(n-1)\cdot n^2}}$$

Eksempel 20 I det tidligere omtalte tyske Lotto er der 49 kugler, og kuglerne er nummereret med tallene $1,2,\ldots,49$. En trækning af de seks kugler foregår netop ved, at kuglerne ikke lægges tilbage, når de først er udtrukket, dvs. vi kan beskrive situationen ved en hypergeometrisk model.

Vi opstiller nu en hypergeometrisk model for den stokastiske variabel X, der tæller antallet af rigtige gæt set i forhold til de seks udtrukne kugler. I den hypergeometriske model er $n=49$, $k=6$ og $a=6$. Sandsynlighederne for den stokastiske variabel beregnes ved:

$$P(X=x)=\frac{K(6,x)\cdot K(49-6,6-x)}{K(49,6)}.$$

x	0	1	2	3	4	5	6
$P(X=x)$	$\frac{K(6,0)\cdot K(49-6,6-0)}{K(49,6)}$						$\frac{K(6,6)\cdot K(49-6,6-6)}{K(49,6)}$

Opgave 36 Udfyld de resterende felter i sandsynlighedstabellen for den hypergeometriske fordeling $h(6,6,49)$ fra eksemplet ovenfor.

Det er ikke alle værktøjsprogrammer, der har den hypergeometriske fordeling indbygget. Hvis dit værktøjsprogram ikke har det, så kan det være en fordel at definere sandsynlighedsfunktionen i dit værkstøjsprogram, så den ikke skal indtastes mange gange. Sandsynligheder for stokastiske variable beregnes herefter blot ved at indsætte den pågældende værdi for den stokastiske variabel.

Når kumulerede sandsynligheder for den stokastiske variabel skal beregnes, kan man bruge sumtegn, så man slipper for at indtaste mange led:

$$P(x_i)+P(x_{i+1})+\ldots+P(x_{j-1})+P(x_j)=\sum_{k=i}^{j}P(x_k).$$

Opgave 37 Ved en tombola med i alt 1000 lodder er der gevinst på halvdelen af lodderne. En person køber 10 lodder.

a) Opstil en hypergeometrisk model for antallet af gevinster blandt personens 10 lodder.

b) Bestem middelværdi og spredning, og fortolk disse tal.

c) Bestem sandsynligheden for ikke at opnå gevinst.

d) Bestem sandsynligheden for at opnå mindst 8 gevinster.

Vi ser nu nærmere på forskellen mellem binomialfordelingen og den hypergeometriske fordeling. Som før nævnt, så benyttes den hypergeometriske fordeling, når en stikprøve udtrækkes uden tilbagelægning, mens binomialfordelingen benyttes når en stikprøve udtrækkes med tilbagelægning. For at illustrere dette, ser vi i den følgende opgave igen på det tyske Lotto, hvor der trækkes seks kugler blandt 49, men hvor vi nu i stedet lader trækningen foregå *med* tilbagelægning.

Opgave 38 Lad den stokastiske variabel X tælle antallet af rigtige gæt i forhold til seks udtrukne kugler i det tyske Lotto, hvor trækningen af seks kugler blandt 49 kugler foregår *med* tilbagelægning. Vi antager nu, at X er binomialfordelt med $n=6$ og $p=\frac{6}{49}$.

a) Udfyld sandsynlighedstabellen:

x	0	1	2	3	4	5	6
$P(X=x)$							

b) Sammenlign dine svar med sandsynlighedstabellen fra eksempel 8.

c) Hvilken af modellerne giver størst sandsynlighed for at opnå 6 rigtige kugler?

I situationer, hvor en stikprøve udtrækkes fra en meget stor population giver binomialfordelingen og den hypergeometriske fordeling stort set samme sandsynlighedsfordeling. Det skyldes, at sandsynligheden for at trække en "succes" stort set ikke ændres af, om de forrige træk er lagt tilbage eller ej. Dette gør sig bl.a. gældende i meningsmålinger, hvor der typisk trækkes en stikprøve i størrelsesorden 1000-2000 personer blandt en alle stemmeberettigede, som er ca. 4 mio. I disse situationer benyttes ofte binomialfordelingen, selvom stikprøven udtrækkes uden tilbagelægning (analyseinstitutter ringer ikke til de samme personer flere gange).

2. Bevis for middelværdi om binomialfordeling

For at bevise sætning 3 tager vi udgangspunkt i binomialformlen: $(p+q)^n = \sum_{k=0}^{n} \binom{n}{k} \cdot p^k \cdot q^{n-k}$.

Sumtegnet $\sum_{k=0}^{n}$ betyder at vi først sætter $k=0$. Herefter sættes $k=1$, så $k=2$ osv. og til sidst $k=n$. Alle disse $n+1$ led lægges sammen. Herved får vi:

$$(p+q)^n = \binom{n}{0} \cdot p^0 \cdot q^n + \binom{n}{1} \cdot p^1 \cdot q^{n-1} + \binom{n}{2} \cdot p^2 \cdot q^{n-2} + \ldots + \binom{n}{n} \cdot p^n \cdot q^0$$

Opgave 39 Bevis sætning 3 ved at løse nedenstående opgaver.

Beviset tager som nævnt udgangspunkt i binomialformlen, som vi i det følgende opfatter som en funktion af p.

a) Vis, at man ved at differentiere venstre side af binomialformlen med hensyn til p får

$$n \cdot (p+q)^{n-1}.$$

b) Vis, at man ved at differentiere højre side af binomialformlen med hensyn til p får

$$\sum_{k=1}^{n} \binom{n}{k} \cdot k \cdot p^{k-1} \cdot q^{n-k}.$$

c) Sæt de to udtryk for differentialkvotienten lig hinanden, og gang ligningen igennem med p. Vis, at man hermed får nedenstående.

$$n \cdot p \cdot (p+q)^{n-1} = \sum_{k=1}^{n} \binom{n}{k} \cdot k \cdot p^k \cdot q^{n-k} = \sum_{k=0}^{n} \binom{n}{k} \cdot k \cdot p^k \cdot q^{n-k}.$$

Bemærk, at der i sidste lighedstegn skiftes fra at summere fra $k=0$ til $k=1$. Hvorfor kan dette ekstra led uden videre tilføjes her?

d) Ovenstående formel gælder for alle p, q og n, og den gælder derfor også i det tilfælde, hvor $q=1-p$ (dvs. når $p+q=1$). Indsæt dette i formlen og reducér til:

$$n \cdot p = \sum_{k=0}^{n} \binom{n}{k} \cdot k \cdot p^k \cdot (1-p)^{n-k}.$$

e) Skriv nu højre side ud. Da $p_k = P(X=k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$, kan ligningen skrives som $n \cdot p = 0 \cdot p_0 + 1 \cdot p_1 + 2 \cdot p_2 + \ldots + n \cdot p_n$.

f) Sammenlign højre side med definitionen af middelværdien (definition 3), og færdiggør beviset.