Sandsynligheds-regning og statistik

B-niveau STX

Lille opdatering 27.08.20.
Lille opdatering 12.09.20. Forkert nummerering af opgaver i kapitel 5 er rettet + ekstra opgave.
Opdatering i Appendiks A med mere, 24.09.20.
Opdatering i afsnit 4.5, 23.03.21.

Kontakt: vestergaard@matematiksider.dk

Forsidebillede: Haderslev Domkirke set fra den indre dam.

Indholdsfortegnelse

Forord

Denne lille e-bog er skrevet til at kunne dække det meste af pensum i sandsynlighedsregning og statistik i gymnasiets B-niveau, STX. Deskriptiv statistik er dog ikke medtaget. Du kan finde en note andetsteds på min hjemmeside angående dette emne. Hvad angår A-niveau, STX, så skal de have lidt mere statistik. Den smule ekstrastof, som vedrører regressionsanalyse, er dækket ind i det ekstra afsnit 5.5 om konfidensintervaller for hældning. Hvad angår sandsynlighedsregning, så mangler der noget mere med normalfordelingen til A-niveau. Det vil der blive lavet en anden lille note om.

Erik Vestergaard
Haderslev Katedralskole

1. Kombinatorik

1.1 Multiplikationsprincippet og additionsprincippet

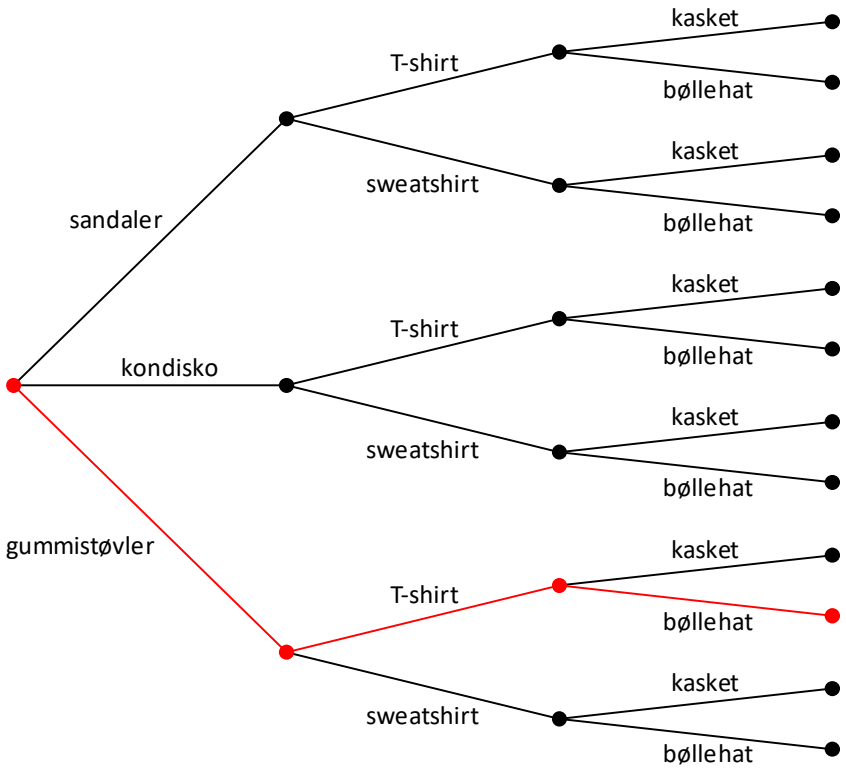
Ifølge Gyldendals *Den Store Danske* er *kombinatorik* kunsten at tælle endelige mængder. I dette kapitel skal vi se på nogle teknikker til at foretage optællinger. Det er særlig vigtigt i forbindelse med udregning af sandsynligheder. Sidstnævnte kommer vi dog først til i næste kapitel.

Sætning 1.1 (Multiplikationsprincippet – "både og")

Antag at et *kombineret valg* består af en række delvalg foretaget i ordnet rækkefølge 1, 2, … k. Antag desuden, at der i forbindelse med det i'te valg er n_i valgmuligheder, uanset hvilke valg, der blev foretaget i de forrige $i-1$ delvalg. Da er der for det kombinerede valg samlet set $n_1 \cdot n_2 \cdot \ldots \cdot n_k$ valgmuligheder.

Eksempel 1.2

Børnene i en børnehave skal klædes på til en udflugt. Børnene kan foruden deres egne bukser vælge mellem følgende, som børnehaven stiller frit til rådighed: fodtøj (*sandaler*, *kondisko* eller *støvler*), overdel (*T-shirt* eller *sweatshirt*) og hat (*kasket* eller *bøllehat*). Antallet af måder, hvorpå børnene kan foretage det kombinerede valg af børnehavens beklædningsgenstande er derfor ifølge multiplikationsprincippet $3 \cdot 2 \cdot 2 = 12$. Det kan illustreres ved et såkaldt *tælletræ*.



Et kombineret valg kan for eksempel være det, der på figuren er markeret med rødt: Gummistøvler, T-shirt og bøllehat.

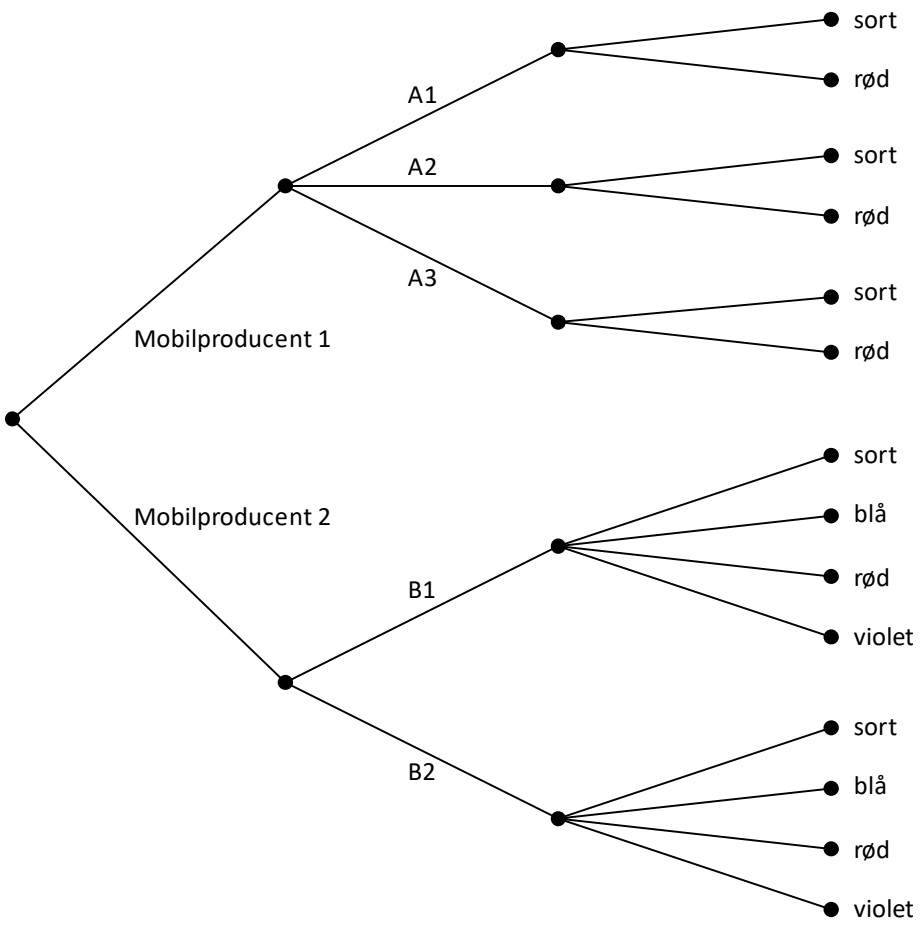
□

Sætning 1.3 (Additionsprincippet – "enten eller")

Hvis man i en valgsituation kommer ud for at skulle vælge mellem én ud af *enten* n_1 muligheder *eller* én ud af n_2 muligheder, *eller* én ud af n_3 muligheder, etc., da bliver det totale antal valgmuligheder $n_1 + n_2 + \ldots + n_k$.

Eksempel 1.4

Louise skal have ny mobiltelefon. Hun har bestemt sig til at vælge én blandt to forskellige brands. Mobilproducent 1 har 3 modeller: A1, A2 og A3. Hver af disse modeller kommer i to forskellige farver: sort eller rød. Mobilproducent 2 derimod kommer kun i to forskellige modeller B1 og B2. Hver af modellerne fås til gengæld i fire forskellige farver: Sort, blå, rød og violet. Spørgsmålet er, hvor mange forskellige mobiltelefoner, Louise kan vælge imellem? Man kan eventuelt stille det op i et tælletræ.



Vi kan ikke udelukkende bruge multiplikationsprincippet, da der ikke er lige mange valgmuligheder i hver kategori. Vi kan imidlertid dele problemet op i to: Først finder vi, hvor mange valgmuligheder der er, hvis man vælger henholdsvis mobilproducent 1 og mobilproducent 2, og derefter lægger man de to tal sammen (additionsprincippet: enten-eller). I tilfældet med mobilproducent 1, kan vi benytte multiplikationsprincippet: her er $3 \cdot 2 = 6$ valgmuligheder. I tilfældet med mobilproducent 2 er der tilsvarende $2 \cdot 4 = 8$ muligheder. Til sidst får vi ved brug af additionsprincippet antallet af valgmuligheder til:

$$3 \cdot 2 + 2 \cdot 4 \; = \; 6 + 8 \; = \; 14$$

□

Bemærkning 1.5

I visse situationer kan det være ganske svært at tælle antallet af muligheder korrekt op. Opdelingen af et problem og senere brug af additionsprincippet er ofte meget brugbar, men man skal være varsom med, at man får talt alle muligheder med én og kun én gang. Det kan man for eksempel se i eksempel 1.14 b) længere fremme.

1.2 Permutationer og kombinationer

Et af nøgleredskaberne i kombinatorikken er følgende:

Definition 1.6

For ethvert $n \in N$ defineres tallet $n!$ ved:

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

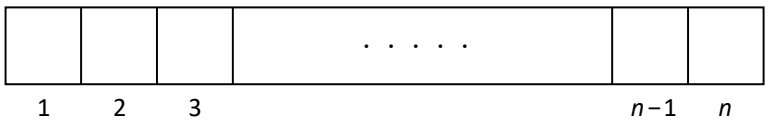
Specielt defineres $0! \; = \; 1$. Tallet $n!$ betegnes "n fakultet" eller bare "n udråbstegn".

Dets betydning skyldes især følgende sætning, som er en grundsten i mange optællinger:

Sætning 1.7

Antallet af måder, hvorpå man kan bytte rundt på rækkefølgen af n forskellige elementer, er $n!$

Bevis: Nedenfor er vist en liste med n tomme pladser, nummereret fra 1 til og med n.



På første plads kan anbringes n forskellige elementer. Da der er $n-1$ elementer tilbage, kan den næste plads besættes på $n-1$ forskellige måder. Da der er $n-2$ elementer tilba-

ge, kan den tredje plads besættes på $n-2$ forskellige måder, etc. indtil den sidste plads, hvor der kun er ét element at vælge. Ifølge multiplikationsprincippet fås i alt

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

muligheder for at placere de n forskellige elementer.

□

At bytte rundt på rækkefølgen af elementer kalder man med et matematisk udtryk at *permutere* elementerne. En *permutation* af en række elementer er altså en ombytning af den rækkefølge, som elementernes står i.

Eksempel 1.8

Bogstaverne a, b, c og d kan ifølge sætning 1.7 permuteres på $4!=1\cdot 2\cdot 3\cdot 4=24$ måder. Man kan opskrive de mulige permutationer på følgende måde:

a b c d	b a c d	c a b d	d a b c
a b d c	b a d c	c a d b	d a c b
a c b d	b c a d	c b a d	d b a c
a c d b	b c d a	c b d a	d b c a
a d b c	b d a c	c d a b	d c a b
a d c b	b d c a	c d b a	d c b a

Man kan nemt miste overblikket, hvis ikke man er systematisk. I søjle 1 står alle permutationer, der starter med a, i anden søjle dem, der starter med b, etc.

□

Eksempel 1.9

I et klasseværelse er der 20 stole. Antallet af forskellige måder, hvorpå en klasse med 20 elever kan indtage pladserne er:

$$20! \;=\; 1\cdot 2\cdot 3\cdot\ldots\cdot 20 \;=\; 2432902008176640000$$

altså et meget overvældende antal måder.

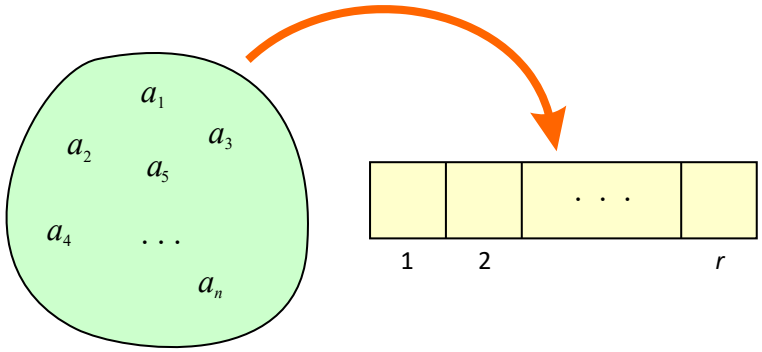
□

Sætning 1.10 (Permutationer)

Lad r og n være naturlige tal med $r \le n$. Antallet af måder at udtrække r elementer ud af en mængde med n forskellige elementer regnet *med rækkefølge* er givet ved:

$$P(n,r) = \frac{n!}{(n-r)!} \tag{2}$$

Bevis: Vi har en mængde bestående af n forskellige elementer $a_1, a_2, \ldots, a_n$. Vi udvælger r af dem og anbringer dem på de r første felter i listen nedenfor. Rækkefølgen af disse betyder noget her.



Den første plads kan besættes på n forskellige måder. Plads nummer 2 kan besættes på $n-1$ forskellige måder, etc. indtil den r'te plads, som kan besættes på $n-r+1$ forskellige måder. Ifølge multiplikationsprincippet giver det følgende antal muligheder:

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

hvor vi, for at få et mere kompakt udtryk, har forlænget med $(n-r)\cdot\ldots\cdot 2\cdot 1$ i tæller og nævner. Det ønskede er dermed vist.

□

Sætning 1.11 (Kombinationer)

Lad r og n være naturlige tal med $r \le n$. Antallet af måder at udtrække r elementer ud af en mængde med n forskellige elementer regnet *uden rækkefølge* er givet ved:

$$K(n,r) = \frac{n!}{r!(n-r)!} \tag{3}$$

Bevis: Vi kan få resultatet ved at tage udgangspunkt i resultatet fra sætning 1.10. Nu er vi ikke længere interesseret i rækkefølgen af de r elementer. Derfor vil enhver permutation af de samme r elementer fra forrige sætning resultere i den samme udtrækning. Antallet af måder, vi kan permutere de samme r elementer, er ifølge sætning 1.7 lig med $r!$. Derfor fås antal muligheder i den nye situation uden rækkefølge ved at dividere antallet af muligheder fra sætning 1.10 med $r!$ Heraf fås det ønskede. □

Især den sidste formel med *kombinationer* er én, man kommer til at bruge rigtig meget i kombinatorikken og dermed også sandsynlighedsregningen. Det skyldes, at der er tale om en meget generel situation, som optræder i mange praktiske anvendelser. Undertiden benyttes der en lidt anden notation for $K(n,r)$:

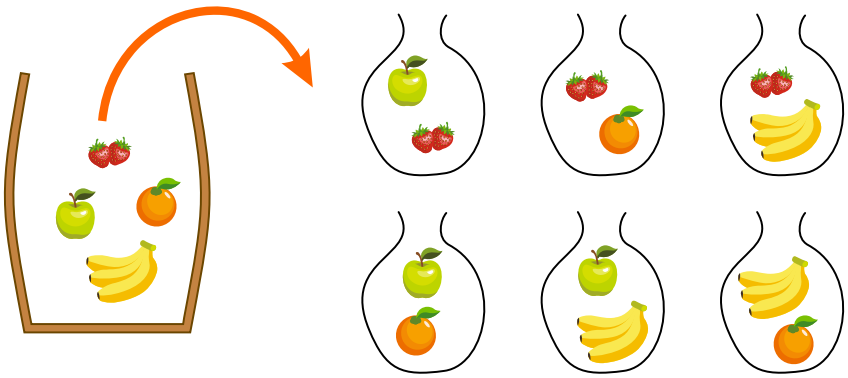
(4) $$K(n,r) = K_{n,r} = \binom{n}{r}$$

Den sidste skrivemåde med parenteser udtales "n over r".

Eksempel 1.12

Lad os sige, at man ønsker at udtrække to frugter ud af en tønde med fire forskellige frugter. Det kan ifølge sætning 1.11 gøres på seks forskellige måder:

$$K(4,2) = \frac{4!}{2!(4-2)!} = \frac{4!}{2!\cdot 2!} = \frac{1\cdot 2\cdot 3\cdot 4}{1\cdot 2\cdot 1\cdot 2} = 6$$



□

Eksempel 1.13

En fodboldtræner har rådighed over 20 spillere og skal sætte et hold på 11 spillere.

a) På hvor mange måder kan han vælge de spillere 11, som skal i kamp?
b) På hvor mange måder kan han sætte holdet på, når man tager hensyn til spillernes plads på holdet (angriber, venstre back, etc.)?

Løsning:

a) Her er det klart, at rækkefølgen er ligegyldig, så vi benytter sætning 1.11:

$$\binom{20}{11} = \frac{20!}{11!\,(20-11)!} = \frac{20!}{11!\,9!} = 167960$$

b) I dette tilfælde er rækkefølgen ikke ligegyldig, så vi skal benytte sætning 1.10, og får et langt større antal muligheder:

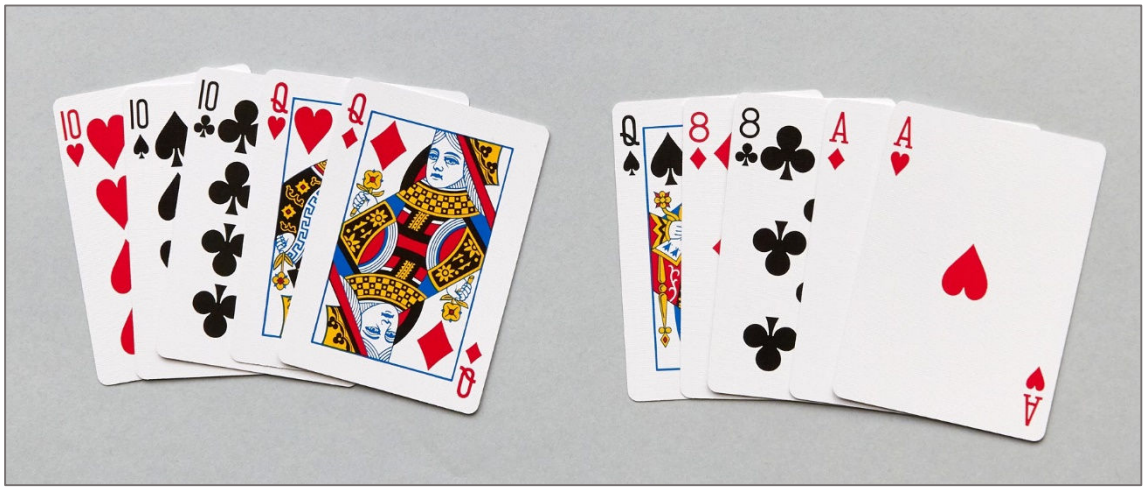
$$P(20,11) = \frac{20!}{(20-11)!} = \frac{20!}{9!} = 6704425728000$$

Det er klart, at man i denne opgave bør bruge CAS-værktøjets funktioner! □

Eksempel 1.14 (Kortspil)

I poker kaldes det for *fuldt hus*, hvis man får tre ens og to ens. Der uddeles fem kort fra et kortspil med 52 kort (ingen jokere).

a) På hvor mange måder kan man få fuldt hus?
b) På hvor mange måder kan man få to par?



Løsning:

a) Man kan tænke således: Først vælges den talværdi, som de tre ens skal have. Det svarer til at udvælge 1 tal ud af 13, som kan gøres på $K(13,1)$ måder. Dernæst vælger man de tre kulører, som de tre ens skal have. Da der er fire kulører, kan det gøres på $K(4,3)$ måder. Dernæst skal vi have valgt talværdien på de to ens. Det kan gøres på $K(12,1)$ måder, da den ene talværdi jo er gået. Endelig skal man vælge hvilke kulører, de to ens skal have. Her er der $K(4,2)$ muligheder. Vi bruger multiplikationsprincippet for at få det samlede antal kombinationer:

$$K(13,1)\cdot K(4,3)\cdot K(12,1)\cdot K(4,2) \;=\; 3744$$

b) Situationen her er lumsk, for hvis man uden eftertanke imiterer proceduren fra a), så kunne man komme frem til $K(13,1)\cdot K(4,2)\cdot K(12,1)\cdot K(4,2)\cdot K(11,1)\cdot K(4,1)$, som giver 247104. Men det er forkert! Problemet er, at man har talt hver kombination med dobbelt! Det kunne jo være, at man i nævnte rækkefølge valgte 8 i klør og ruder og derefter 13 (es) i hjerter og ruder, og til sidst 12 (dame) i spar. Men dette er den samme hånd, som man får ved først at vælge 13 (es) i hjerter og ruder og derefter 8 i klør og ruder og til sidst 12 (dame) i spar. Den rigtige fremgangsmåde er at tænke således: Først udtages de talværdier, som de to par skal have. Det kan gøres på i alt $K(13,2)$ måder. Dernæst vælger vi kulørerne for den højeste talværdi og derefter kulørerne for den laveste talværdi. Det kan i begge tilfælde gøres på $K(4,2)$ måder. Endelig skal vi have taget et kort med en talværdi, som *ikke* allerede er taget. Det giver $K(11,1)$ måder. Kuløren på dette kort kan vælges på $K(4,1)$ måder. I alt fås:

$$K(13,2)\cdot K(4,2)\cdot K(4,2)\cdot K(11,1)\cdot K(4,1) \;=\; 123552$$

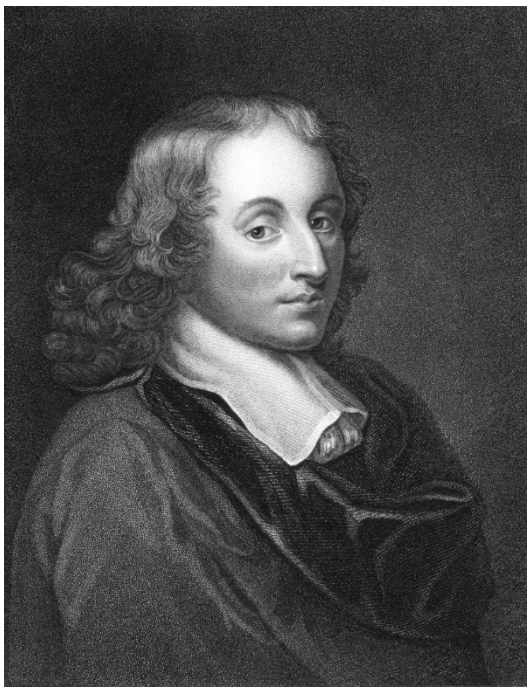
Det er det rigtige antal kombinationer med 2 par.

2. Endeligt sandsynlighedsfelt

2.1 Kort historisk

Allerede i 1500-tallet regnede den italienske matematiker og videnskabsmand *Gerolamo Cardano* (1501-1576) på forskellige odds og sandsynligheder i forbindelse med terningekast. Hans resultater blev dog ikke publiceret mens han levede. Selv var han var en af de mest indflydelsesrige matematikere under renæssancen. Blandt andet er han også kendt for at finde løsninger til 3. gradsligninger. Cardonos undersøgelser indenfor sandsynlighedsregningen strakte sig ret langt, men hans og andre italienske matematikeres idéer gik efterhånden i glemmebogen.

Det var først i midten af 1600-tallet, at man for alvor så nye tiltag på området. Her startede det med, at en ivrig hasard-spiller, *Chevalier de Méré*, spurgte den franske matematiker, fysiker og opfinder *Blaise Pascal* (1623-1662) til råds i et spørgsmål om nogle spilsituationer. Det affødte blandt andet en korrespondance mellem Pascal og matematikeren og juristen *Pierre de Fermat* (1607-1665). I fællesskab med Fermat grundlagde Pascal efterfølgende den klassiske sandsynlighedsregning.



Blaise Pascal (1623-1662)

Pierre-Simon Laplace (1749-1827)

I det efterfølgende forløb ses matematikeren og videnskabsmanden *Christiaan Huygens* (1629-1695) samt matematikerne *Jacob Bernoulli* (1655-1705) og *Abraham de Moivre* (1667-1754) at yde vigtige bidrag til sandsynlighedsregningen. Det næste kæmpe bidrag stod franskmanden *Pierre-Simon Laplace* (1749-1827) dog for. Han gjorde sig gældende indenfor ingeniørvidenskab, matematik, statistik, fysik, astronomi og filosofi og bliver undertiden omtalt som den franske udgave af Newton. Han leverede utallige resultater i sandsynlighedsregningen, med *den centrale grænseværdisætning* som kronen på værket. Denne sætning, som det her vil være for kompliceret at gengive i sin ordlyd, kobler på forunderlig vis nogle helt centrale og generelle aspekter ved stikprøver med den såkaldte

normalfordeling. Sidstnævnte var den berømte tyske matematiker *Carl Friedrich Gauss* (1777-1855) blandt andet med til at indføre. Med normalfordelingen havde man altså ramt noget vigtigt og på en vis måde universelt. Mange andre sandsynlighedsfordelinger blev studeret, men normalfordelingen forbliver den mest anvendte. I nyere tid var det russeren *Andrey Kolmogorov* (1903-1987), som var med til at systematisere og aksiomatisere sandsynlighedsregningen til den velpolerede og logiske form, vi ser i dag. I dagens samfund spiller sandsynlighedsregning og statistik en overordentlig stor rolle. Det handler ikke blot om at udregne sandsynligheder i forbindelse med spil på casinoer. Indenfor økonomi og forsikringsbranchen opstiller man stokastiske modeller for at blive klogere på, hvordan forskellige mekanismer fungerer i samfundet. Medicinalfirmaerne har statistikere ansat til at bearbejde data fra kliniske tests med henblik på at godtgøre et produkts virkning på patienterne. Med computernes store regnekraft og Internettets store udbredelse er begreber som Big Data desuden kommet på banen.

2.2 Endeligt sandsynlighedsfelt

Historisk set har beregning af sandsynligheder givet anledning til mange misforståelser og fejltagelser. Heldigvis er sandsynlighedsregningen blevet systematiseret med vigtige begreber, så det nu er klart, hvad der tales om. Først skal vi definere det, der kaldes et *endeligt sandsynlighedsfelt*. Læseren opfordres til at gøre sig begreberne klart!

Definition 2.1

Et *endeligt sandsynlighedsfelt* (U, P) består af en endelig mængde $U = \{u_1, u_2, \ldots, u_n\}$ kaldet *udfaldsrummet,* samt en sandsynlighedsfunktion P defineret på U, opfyldende:

a) $0 \le P(u) \le 1$ for alle $u \in U$

b) $P(u_1) + P(u_2) + \ldots + P(u_n) = 1$

Sidstnævnte skriver vi ofte lidt kortere med et sumtegn: $\sum_{i=1}^{n} P(u_i) = 1$.

$P(u_i)$ kaldes for sandsynligheden for udfaldet u_i.

Eksempel 2.2

Ved kast med en terning, hvor vi betragter det antal øjne, som vender opad, har vi altså udfaldsrummet $U = \{1, 2, 3, 4, 5, 6\}$. Hvis der er tale om en ægte terning, kan vi umiddelbart (*a priori*) sige, at $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$. Ofte vil vi angive *sandsynlighedsfordelingen* i en tabel:

u	1	2	3	4	5	6
$P(u)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Vi ser klart, at betingelserne for et endeligt sandsynlighedsfelt fra definition 2.1 er opfyldt: Dels er alle sandsynlighederne tal mellem 0 og 1, og så giver de 1 tilsammen:

$$\tfrac{1}{6}+\tfrac{1}{6}+\tfrac{1}{6}+\tfrac{1}{6}+\tfrac{1}{6}+\tfrac{1}{6}=1$$

□

Eksempel 2.3

Der fødes et barn. Udfaldsrummet er $U=\{\text{dreng, pige}\}$. Det oplyses, at sandsynligheden for at få en dreng er 51%. Af egenskab b) for et endeligt sandsynlighedsfelt kan vi derfor selv udregne den sidste sandsynlighed: $P(\text{pige})=1-P(\text{dreng})=1-0{,}515=0{,}485$.

u	dreng	pige
$P(u)$	0,51	0,49

□

Bemærkning 2.4

Eksempel 2.2 er et eksempel på et endeligt sandsynlighedsfelt, hvor sandsynlighederne er givet *a priori*. Erfaringen giver os uden behov for eksperimenter, at hvert udfald har sandsynligheden 1/6. Derimod beror sandsynlighederne i eksempel 2.3 på erfaringen efter optællingen ved en lang række fødsler. Her viser det sig, at ca. 51,5% af fødslerne resulterer i en dreng. Tallet er dog ikke helt stabilt, som man kan læse om i en artikel fra Illustreret Videnskab fra 1/9-2019. Det afhænger af flere faktorer. En sandsynlighed, som baserer sig på statistisk data som i eksempel 2.3, betegnes undertiden en *frekventiel* sandsynlighed – stammende fra ordet *frekvens*.

Definition 2.5

En delmængde H af et udfaldsrum U kaldes en *hændelse*. Vi skriver: $H\subseteq U$. Tallet

$$P(H)=\sum_{u\in H}P(u) \tag{1}$$

kaldes *sandsynligheden for hændelsen H*. Specielt sætter vi $P(\varnothing)=0$, hvor $\varnothing$ er den tomme mængde. Specielt gælder $P(U)=1$.

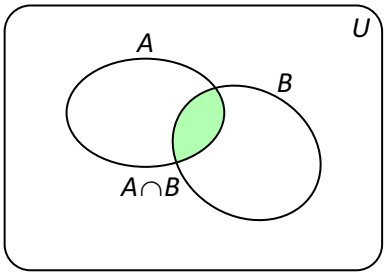
Eksempel 2.6

I forbindelse med eksempel 2.2, hvor der udføres kast med én terning: Betragt hændelsen H: "Terningen viser mindst 5". Som mængde betragtet er hændelsen altså $H=\{5,6\}$. Definition 2.5 giver da: $P(H)=P(5)+P(6)=\tfrac{1}{6}+\tfrac{1}{6}=\tfrac{1}{3}$. Så sandsynligheden for, at man får mindst en femmer ved terningekastet er altså 1/3.

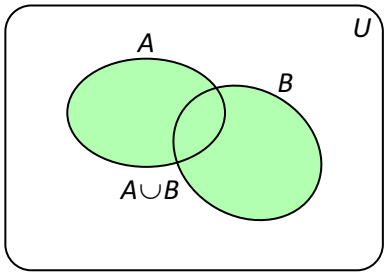
□

Vi skal kigge på forskellige operationer, man kan foretage på mængder. Først definerer vi operationerne og derefter illustreres med såkaldte *Venn diagrammer*.

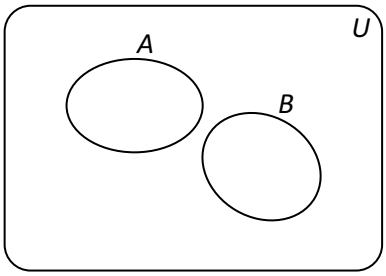
Fællesmængde	$A \cap B$ består af de elementer, som er i både A og B.
Foreningsmængde	$A \cup B$ består af de elementer, som er i A og/eller i B.
Disjunkte mængder	A og B kaldes disjunkte, hvis A og B ikke har nogen elementer til fælles, dvs. hvis $A \cap B = \varnothing$.
Komplementærmængde	A^c består af alle de elementer, som er i U, men ikke i A.
Delmængde	$B \subseteq A$ hvis ethvert element fra B også er i A. Det kan alternativt udtrykkes ved at $u \in B \Rightarrow u \in A$.
Klassedeling	Mængderne $A_1, A_2, \ldots, A_n$ kaldes en klassedeling af A, hvis mængderne to og to er indbyrdes disjunkte, og foreningsmængden af dem alle er lig med B. Det kan også udtrykkes ved: $A_i \cap A_j = \varnothing$ for alle $i \neq j$ og $A_1 \cup A_2 \cup \ldots \cup A_n = A$.



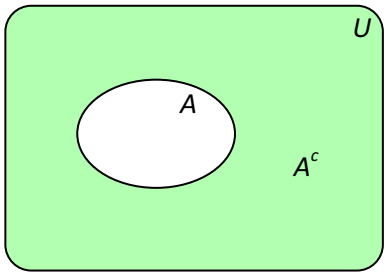
Fællesmængde



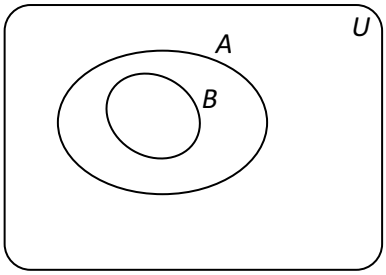
Foreningsmængde



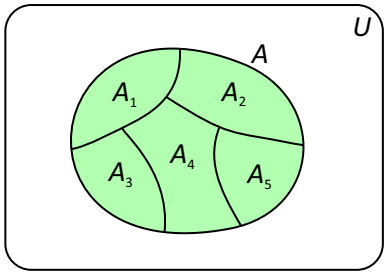
Disjunkte mængder



Komplementærmængde



Delmængde



Klassedeling

Sætning 2.7

For hændelser i samme sandsynlighedsfelt gælder:

a) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

b) For disjunkte hændelser A og B gælder: $P(A \cup B) = P(A) + P(B)$

c) $P(A^c) = 1 - P(A)$

d) For en klassedeling $A_1, A_2, \ldots, A_n$ af A gælder:

$$\sum_{i=1}^{n} P(A_i) = P(A_1) + P(A_2) + \ldots + P(A_n) = P(A)$$

Bevis: a) Vi ved at sandsynligheden for en hændelse fås ved at addere sandsynlighederne af de enkelte udfald i hændelsen. Når man beregner summen $P(A) + P(B)$ bliver sandsynlighederne for udfaldene i $A \cap B$ talt med to gange. Derfor skal man trække sandsynligheden af $A \cap B$ fra, for at få sandsynligheden for $A \cup B$. Det overlades til læseren at bevise de øvrige punkter.

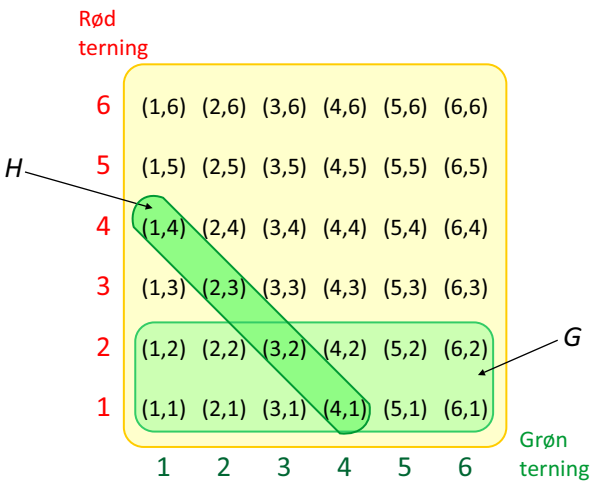
□

Eksempel 2.8

I eksempel 2.2 og 2.6 kiggede vi på kast med en terning. Nu vil vi studere kast med to terninger. For at gøre tingene mere tydelige vedtager vi, at den ene terning er grøn og den anden rød. En måde at beskrive et udfald ved et kast med de to terninger er som et koordinatsæt. Udfaldet $(2,4)$ står således for, at den grønne terning viste 2, mens den røde viste 4 øjne. Udfaldsrummet består således af 36 mulige udfald:

$$U = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(2,2),\ldots,(6,6)\}$$

For at få større overblik i det følgende, kan det være hensigtsmæssigt at anbringe de 36 udfald i et kvadrat. Vi skal nemlig studere nogle hændelser på næste side.



Lad H være den hændelse, som med ord kan udtrykkes: "Summen af terningerne er 5". Som mængde kan den skrives:

$$H = \{(1,4),(2,3),(3,2),(4,1)\}$$

En anden hændelse er G: "Den røde terning viser højst 2". Som mængde kan den skrives:

$$G = \{(1,1),(2,1),(3,1),(4,1),(5,1),(6,1),(1,2),(2,2),(3,2),(4,2),(5,2),(6,2)\}$$

Mængderne er angivet lidt mere overskueligt som delmængder i kvadratet på figuren på forrige side. Vi kan foretage mængdeoperationer, fx tage fællesmængden:

$$H \cap G = \{(3,2),(4,1)\}$$

Den består af de to udfald, som ligger i begge delmængder. Hændelsen kan med ord beskrives: "Summen af terningerne er 5 **og** den røde terning viser højst 2". Vi kan også vælge at tage foreningsmængden:

$$H \cup G = \{(1,1),(2,1),(3,1),(4,1),(5,1),(6,1),(1,2),(2,2),(3,2),(4,2),(5,2),(6,2),(1,4),(2,3)\}$$

Foreningsmængden består af de 14 udfald, som ligger i enten den ene mængde eller i den anden mængde. Med ord kan hændelsen beskrives: "Summen af terningerne er 5 **eller** den røde terning viser højst 2".

Indtil nu har vi udelukkende talt om mængder eller hændelser. Vi skal vi til at se på sandsynlighederne for de enkelte udfald. For det første bør det stå læseren klart, at alle de 36 udfald i udfaldsrummet U har den samme sandsynligheder for at forekomme. Derfor er sandsynligheden for hver eneste udfald i U lige med $\frac{1}{36}$. Det giver anledning til følgende sandsynligheder:

$$P(H) = 4\cdot\tfrac{1}{36} = \tfrac{4}{36} = \tfrac{1}{12}, \quad P(G) = 12\cdot\tfrac{1}{36} = \tfrac{12}{36} = \tfrac{1}{3},$$
$$P(H\cap G) = 2\cdot\tfrac{1}{36} = \tfrac{2}{36} = \tfrac{1}{18}, \quad P(H\cup G) = 14\cdot\tfrac{1}{36} = \tfrac{14}{36} = \tfrac{7}{18}$$

Nu talte vi op direkte for at bestemme sandsynligheden for foreningsmængden. Vi kunne også have brugt sætning 2.7 a):

$$P(H\cup G) = P(H)+P(G)-P(H\cap G) = \tfrac{4}{36} + \tfrac{12}{36} - \tfrac{2}{36} = \tfrac{14}{36} = \tfrac{7}{18}$$

□

Punkt c) i sætning 2.7 er ofte brugbar. Der er opgaver, hvor man ønsker at udregne sandsynligheden for en hændelse A, men hvor den er besværlig at udregne direkte, hvorimod sandsynligheden for den komplementære hændelse A^c er meget nemmere at udregne.

Eksempel 2.9

Bestem sandsynligheden for at få plat mindst én gang ved fire kast med en mønt. Udfaldene i eksperimentet kan, i stil med terningeforsøgene, passende opskrives som et 4-tuppel. Udfaldet (p,p,k,p) betyder således, at de første to kast gav plat, det tredje kast gav

krone, og det sidste gav plat. Det er oplagt, at der er 16 udfald i udfaldsrummet. Da de i dette tilfælde er lige sandsynlige, har hvert udfald altså sandsynligheden $\frac{1}{16}$. Man kunne begynde at undersøge, hvilke af udfaldene, som ligger i hændelsen A: *Der er mindst én plat*, og derefter addere deres sandsynligheder. Det er imidlertid meget nemmere at betragte den komplementære hændelse A^c: *Alle kast viste krone*. I denne hændelse er der kun udfaldet (k,k,k,k). Sætning 9c) giver nu:

$$P(A) = 1 - P(A^c) = 1 - \frac{1}{16} = \frac{15}{16}$$

□

Nu til en type sandsynlighedsfelt, som man ofte støder på:

Definition 2.10 (Symmetrisk sandsynlighedsfelt)

Et (endeligt) *symmetrisk* sandsynlighedsfelt er ét, hvor alle udfald har samme sandsynlighed.

Vi har allerede arbejdet med symmetriske sandsynlighedsfelter i flere af de eksempler, vi hidtil har gennemgået. For sådanne sandsynlighedsfelter er det særligt nemt at udregne sandsynligheder, som følgende sætning siger:

Sætning 2.11

I et symmetrisk sandsynlighedsfelt gælder følgende for hændelsen H:

(2) $$P(H) = \frac{\text{Antal gunstige udfald}}{\text{Antal mulige udfald}}$$

Eksempel 2.12

I eksempel 2.2 med kast med en terning har vi klart at gøre med et symmetrisk sandsynlighedsfelt. Lad os sige, at vi ønsker at bestemme sandsynligheden for hændelsen H: "Terningen viser mindst 5". Vi ser her, at antal gunstige udfald er 2, svarende til udfaldene 5 og 6. Antal mulige udfald er 6. Dermed fås:

$$P(H) = \frac{\text{Antal gunstige udfald}}{\text{Antal mulige udfald}} = \frac{2}{6} = \frac{1}{3}$$

□

Eksempel 2.13 (Fuldt hus i poker)

Vi skal også se på et mere kompliceret eksempel med et symmetrisk sandsynlighedsfelt, hvor vi vil udnytte kombinatorikken fra forrige kapitel til at udregne antal muligheder. Lad os sige, at vi ønsker at beregne sandsynligheden for at få *fuldt hus* i Poker ved uddeling af 5 kort fra et kortspil med 52 kort – altså uden jokere.

Der er klart tale om et symmetrisk sandsynlighedsfelt, fordi ethvert af de 52 kort er lige sandsynligt at modtage. Allerede i eksempel 1.14 i forrige kapitel udregnede vi antallet af måder, hvorpå man kan få fuldt hus. Vi gentager lige hurtigt argumenterne: Talværdien for kortene med de tre ens kan vælges på $K(13,1)$ måder, de tre kulører ud af de fire mulige for de tre ens kan vælges på $K(4,3)$ måder. Så er der de to ens. Talværdien kan vælges på $K(12,1)$ måder, da den ene talværdi er taget. Kulørerne på de to ens kan vælges på $K(4,2)$ måder. Multiplikationsprincippet giver os:

$$\text{Antal gunstige udfald} \;=\; K(13,1)\cdot K(4,3)\cdot K(12,1)\cdot K(4,2) \;=\; 3744$$

En tilfældig hånd kan vælges på $K(52,5)$ måder, da der skal vælges 5 kort ud af 52.

$$\text{Antal mulige udfald} \;=\; K(52,5) \;=\; 2598960$$

Dermed fås ifølge sætning 2.11:

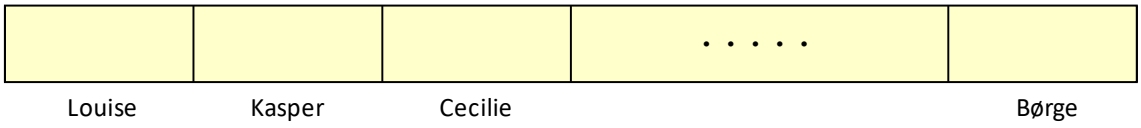
$$P(\text{fuldt hus}) \;=\; \frac{\text{Antal gunstige udfald}}{\text{Antal mulige udfald}} \;=\; \frac{3744}{2598960} \;=\; 0{,}00144$$

Man får altså fuldt hus ca. 1 ud af 700 gange, hvor der uddeles kort – ret sjældent.

□

Eksempel 2.14 (Fødselsdagsparadokset)

Vi skal udregne sandsynligheden for at der i en klasse med 28 elever er mindst to elever, der har fødselsdag på samme dag. Vi kalder hændelsen at mindst to elever har fødselsdag på samme dag for H. Vi gør den antagelse, at alle dage er lige sandsynlige at have fødselsdag på. Dermed har vi at gøre med et symmetrisk sandsynlighedsfelt.



Det er her en stor fordel at kigge på den komplementære hændelse H^c: *Ingen* elever i klassen har fødselsdag på samme dag. Et gunstigt udfald for den komplementære hændelse vil svare til, at vi ud for hver elev ovenfor skal placere 28 *forskellige* datoer. Antallet af måder, hvorpå man kan placere 28 forskellige datoer i skemaet svarer til antallet af måder at udtrække 28 tal ud af tallene 1, 2, … , 365 *med rækkefølge*! Det giver i alt $P(365,28)$ muligheder ifølge sætning 1.10. Antallet af mulige måder er nemt at udregne: Hver plads kan besættes på 365 måder, når der ikke er noget krav om, hvor elevernes

fødselsdage skal ligge. Ifølge multiplikationsprincippet giver det 365^{28} mulige udfald. Vi er nu klar til at regne:

$$P(H) = 1 - P(H^c) = 1 - \frac{\text{Antal gunstige udfald}}{\text{Antal mulige udfald}}$$
$$= 1 - \frac{P(365,28)}{365^{28}} = 1 - \frac{365 \cdot 364 \cdot 363 \cdot \ldots \cdot 338}{365^{28}} = 0{,}654$$

Sandsynligheden for, at der i en klasse med 28 elever er mindst to elever, der har fødselsdag samme dag, er altså ca. 65%! Vi antog, at alle fødselsdage i løbet af året er lige sandsynlige. Det holder nok ikke helt. Men en lidt skæv fordeling af fødselsdage i den danske befolkning vil kun øge sandsynligheden (overvej). Egentligt er der ikke tale om et paradoks i formel forstand, men det kaldes ofte sådan, fordi resultatet strider så meget imod intuitionen.

□

Uafhængige hændelser

De fleste ved, at når man slår flere gange med en terning, så vil sandsynligheden for at slå et givet antal øjne ikke afhænge af udfaldet af de forrige slag. Vi siger at hændelserne er *uafhængige*. For hvert nyt slag vil sandsynligheden for at slå et bestemt antal øjne altid være 1/6. De fleste ved også, at sandsynligheden for at slå to seksere efter hinanden, eller for dens sags skyld først en 3'er og så en 1'er, er $\frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$. Generelt skal vi definere uafhængighed på følgende måde:

Definition 2.15 (Uafhængighed)

To hændelser A og B *i* et endeligt sandsynlighedsfelt siges at være *uafhængige*, hvis

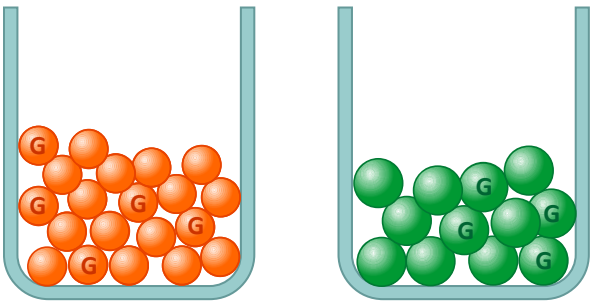
(3) $$P(A \cap B) = P(A) \cdot P(B)$$

Definitionen siger, at to hændelser kaldes uafhængige, hvis sandsynligheden for, at begge hændelser indtræffer, er produktet af sandsynlighederne for de to hændelser.

Eksempel 2.16

I en tombola er der to krukker. Krukke 1 indeholder 20 ens kugler, som man kan lukke op for at se, om der er gevinst. 5 ud af de 20 kugler indeholder en gevinst. I krukke 2 er der 12 kugler, hvoraf de 4 indeholder en gevinst.

a) Der trækkes nu på tilfældig vis en kugle fra hver af de to krukker. Betragt nu følgende hændelser: A: Der var gevinst i krukke 1. B: Der var gevinst i krukke 2.



Krukke 1 Krukke 2

Udfaldsrummet består af alle de par (u_1, u_2), hvor u_1 er en kugle udtrukket fra krukke 1 og hvor u_2 er en kugle udtrukket fra krukke 2. Det er klart, at vi har at gøre med et symmetrisk sandsynlighedsfelt: alle kombinationer er lige sandsynlige. Derfor kan vi bruge sætning 2.11. Vi skal tælle op, hvor mange gunstige og hvor mange mulige udfald, der er for hver hændelse. Lad os se på hændelse A. Antallet af *gunstige* måder at få gevinst i første udtrækning er $5 \cdot 12$, da vi skal have en gevinst i første udtrækning, og det er ligegyldig, hvad der kommer i 2. udtrækning. Antal *mulige* udfald er klart $20 \cdot 12$ ifølge multiplikationsprincippet. Det giver en sandsynlighed på:

$$P(A) = \frac{5 \cdot 12}{20 \cdot 12} = \frac{5}{20} = \tfrac{1}{4}$$

På tilsvarende måde fås:

$$P(B) = \frac{20 \cdot 4}{20 \cdot 12} = \frac{4}{12} = \tfrac{1}{3}$$

Men hvad med fællesmængden af de to hændelser, dvs. $A \cap B$, hvor der skal være gevinst i begge udtrækninger? Antal gunstige muligheder er her klart $5 \cdot 4 = 20$, mens antal mulige stadig er $20 \cdot 12$. Alt i alt:

$$P(A \cap B) = \frac{5 \cdot 4}{20 \cdot 12} = \tfrac{1}{4} \cdot \tfrac{1}{3} = P(A) \cdot P(B)$$

Vi indser, at hændelserne er uafhængige ifølge definition 2.15. Det følger fuldstændig intuitionen: Udfaldet af første udtrækning har ingen som helst indflydelse på udfaldet af anden udtrækning!

b) Nu trækker vi på tilfældigvis måde først en kugle fra krukke 1 og derefter igen en kugle fra krukke 1. Betragt følgende hændelser: A: Der var gevinst ved 1. udtrækning. B: Der var gevinst ved anden udtrækning.

Antallet af gunstige måder at få gevinst i 1. udtrækning er $5 \cdot 19$. Husk at i anden udtrækning er en kugle allerede taget, så der er 19 tilbage. Antal mulige udtrækninger er dermed $20 \cdot 19$. Derfor fås

$$P(A) = \frac{5 \cdot 19}{20 \cdot 19} = \frac{5}{20} = \tfrac{1}{4}$$

Af symmetriårsager fås det samme for hændelse B: $P(B) = \frac{1}{4}$. Fællesmængden af de to hændelser, $A \cap B$, svarer til gevinst i begge udtrækninger. Her er $5 \cdot 4$ gunstige, hvilket giver:

$$P(A\cap B) = \frac{5\cdot 4}{20\cdot 19} = \tfrac{1}{19} \neq P(A)\cdot P(B)$$

Altså er de to hændelser *ikke* uafhængige. Det stemmer også med vores intuition, som fortæller os, at sandsynligheden i 2. udtrækning afhænger af udfaldet af 1. udtrækning. Var der nemlig *ikke* gevinst i første udtrækning, ville der nemlig stadig være 5 kugler med gevinst i til næste udtrækning. Det ville øge sandsynligheden for gevinst i 2. udtrækning i forhold til, hvis der allerede var taget en kugle med gevinst i første udtrækning. Det er her væsentligt, at det er udtrækning *uden tilbagelægning*.

□

Eksempel 2.17

Vi skal se et eksempel, hvor det ikke er helt indlysende, om hændelserne er uafhængige. Betragt følgende to hændelser ved kast med to terninger, ligesom i eksempel 2.8:

A : "Den grønne terning viser 2". B : "Summen af terningerne viser 7".

Hvis man ser på figuren i eksempel 2.8, får man hurtigt ved at tælle gunstige udfald:

$$P(A)=\tfrac{6}{36}=\tfrac{1}{6},\quad P(B)=\tfrac{6}{36}=\tfrac{1}{6},\quad P(A\cap B)=\tfrac{1}{36}$$

Vi ser, at (3) er opfyldt, så hændelserne A og B er uafhængige.

□

Hvad med uafhængighed, hvis der er tre hændelser A, B og C? Hvis man valgte at definere uafhængigheden mellem disse derved, at de parvis skulle være uafhængige via definition 2.15, så viser det sig uheldigvis, at man *ikke* kan sikre den vigtige regel:

(4) $$P(A\cap B\cap C) = P(A)\cdot P(B)\cdot P(C)$$

Derfor vælger man at definere det anderledes. Man vil forlange, at (4) skal gælde, men også $P(A\cap B)=P(A)\cdot P(B)$, $P(A\cap C)=P(A)\cdot P(C)$ og $P(B\cap C)=P(B)\cdot P(C)$ skal være opfyldt. Uafhængighed mellem n hændelser defineres generelt ved:

Definition 2.18

Hændelserne $A_1, A_2, \ldots, A_n$ $(n\geq 2)$ siges at være *uafhængige* (eller *indbyrdes uafhængige*), såfremt der for ethvert udvalg af indices $i_1, i_2, \ldots, i_k$ mellem 1 og n gælder:

$$P(A_{i_1}\cap A_{i_2}\cap\ldots\cap A_{i_k}) = P(A_{i_1})\cdot P(A_{i_2})\cdots P(A_{i_k})$$

Hændelserne i forbindelse med udfaldene af hver terning ved kast med flere terninger er klart uafhængige ifølge denne definition.

2.3 Stokastisk variabel

I afsnit 2.2 indførte vi begrebet et endeligt sandsynlighedsfelt. Man havde givet et udfaldsrum U og en sandsynlighedsfunktion P, som til hvert udfald $u \in U$ knyttede en sandsynlighed $P(u)$, som er et tal mellem 0 og 1. Vi fik også defineret begrebet en hændelse H som værende en delmængde af udfaldsrummet, og sandsynligheden $P(H)$ for hændelsen H blev defineret som summen af sandsynlighederne af alle de udfald, som er i H. I dette afsnit skal vi se på en ofte snedig måde at "udpege" en hændelse på. Ofte er man nemlig interesseret i en hændelse, som har en bestemt egenskab. Det er her det er smart at indføre et nyt begreb: *Stokastisk variabel*.

Definition 2.19 (Stokastisk variabel - diskret)

En *stokastisk variabel* X på et endeligt sandsynlighedsfelt er en funktion fra udfaldsrummet U til mængden af reelle tal R.

Begrebet stokastisk variabel lader sig nemmest forklare ved nogle eksempler.

Eksempel 2.20

Eksperiment: Vi slår med én terning.
Udfaldsrum: $U = \{1, 2, 3, 4, 5, 6\}$
Stokastisk variabel: X er antallet af øjne, som terningen viser.
Det er klart, at X dermed kan antage 6 forskellige værdier. Sandsynlighedsfordelingen for X kan da skrives:

x	1	2	3	4	5	6
$P(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

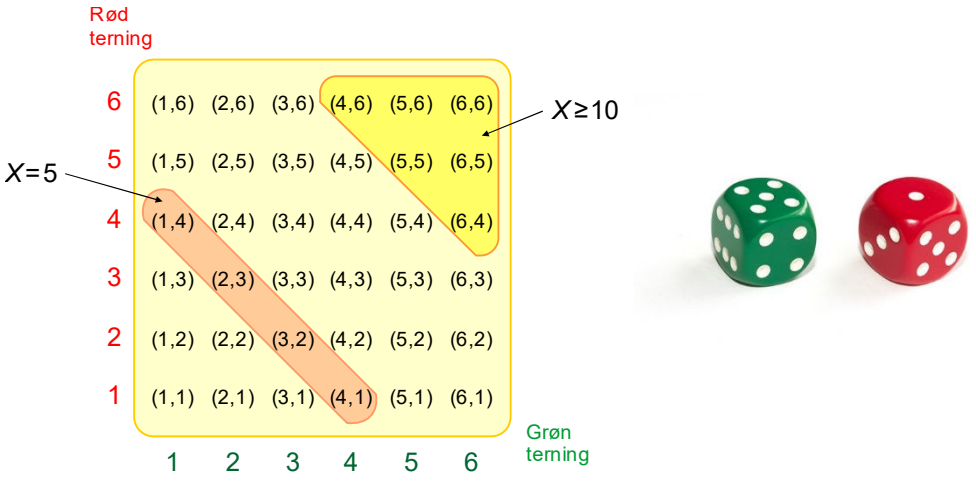
Det bemærkes, at man ikke med sikkerhed kan sige noget om, hvilken værdi den stokastiske variabel antager i eksperimentet. Det er derfor vi kalder den en *stokastisk* variabel, som hentyder til en *tilfældighed*. Vi kan kun udtale os om sandsynlighederne for at X antager nogle givne værdier.

Eksempel 2.21 (Summen af øjnene ved kast med to terninger)

Eksperiment: Vi slår med to terninger – en grøn og en rød.
Udfaldsrum: $U = \{(1,1), (1,2), \ldots, (1,6), (2,1), (2,2), \ldots, (2,6), \ldots, (6,6)\}$
Stokastisk variabel: X er summen af øjnene på de to terninger.
Bemærk, at vi i angivelsen af udfaldsrummet har anvendt samme notation som i eksempel 2.8 fra forrige afsnit. Vi ser, at X her kan antage alle heltallige værdier fra 2 til og med 12. Sandsynlighedsfordelingen kommer til at se således ud:

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

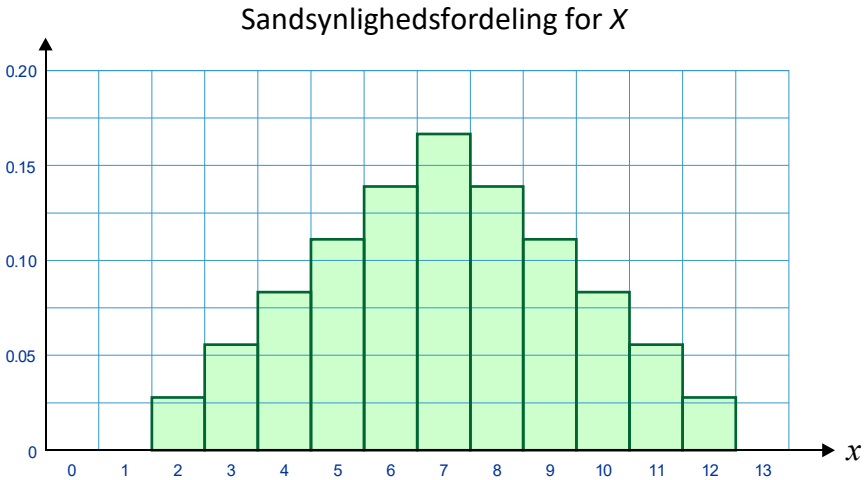
Tankegangen i udregningen af sandsynlighederne er følgende: X antager for eksempel værdien 5 for følgende udfald: $(1,4), (2,3), (3,2)$ og $(4,1)$, jf. skrivemåden i eksempel 2.8 i forrige afsnit.



Når vi skriver $X=5$ mener vi altså hændelsen svarende til den delmængde af udfaldsrummet, som er markeret med orange på figuren ovenfor. Da vi har at gøre med et symmetrisk sandsynlighedsfelt, kan sandsynligheden for hændelsen fås ved at dividere antal gunstige udfald med antal mulige udfald:

$$P(X=5) = \frac{\text{Antal gunstige udfald}}{\text{Antal mulige udfald}} = \frac{4}{36}$$

Sandsynlighederne for de øvrige mulige værdier for x fås på lignende vis. Man kan eventuelt vælge at tegne et pindediagram eller som her et stolpediagram for sandsynlighedsfordelingen for X:



Ikke overraskende er den mest sandsynlige sum af øjnene 7. Man kan også have hændelser givet ved uligheder, fx $X \geq 10$. Hændelsen er markeret på figuren på forrige side. Vi kan se, at der er 6 gunstige udfald. Derfor fås:

$$P(X \geq 10) = \frac{\text{Antal gunstige udfald}}{\text{Antal mulige udfald}} = \frac{6}{36} = \frac{1}{6}$$

I øvrigt har vi også: $P(X \geq 10) = P(X = 10) + P(X = 11) + P(X = 12)$.

□

I næste eksempel skal vi se, at begrebet stokastiske variabel har langt videre perspektiver, end man umiddelbart skulle tro. Igen er eksperimentet kast med to terninger, men den stokastiske variabel angiver nu ikke længere summen af de øjne, terningerne viser. Nu angiver den gevinsten ved et spil!

Eksempel 2.22 (Et spil)

En bankør tilbyder et spil, hvor spilleren slår med to terninger: en grøn og en rød. Hvis der er en 1'er blandt de to terninger, skal spilleren betale 4 kr. til bankøren. I alle andre tilfælde vinder spilleren det beløb i kroner, som svarer til forskellen mellem de to terningers visning. Hvis den ene terning viser 5 og den anden 2, vinder spilleren altså $5 - 2 = 3$ kroner. Lad os være systematiske igen:

Eksperiment: Et kast med to terninger, en grøn og en rød. Antal øjne betragtes.
Udfaldsrum: $U = \{(1,1), (1,2), \ldots, (1,6), (2,1), (2,2), \ldots, (2,6), \ldots, (6,6)\}$, hvor 1. koordinaten i talparret angiver antal øjne for den grønne terning, mens 2. koordinaten angiver antal øjne for den røde terning.
Stokastisk variabel: X angiver det beløb spilleren vinder i ét enkelt spil.

For at bestemme sandsynlighedsfordelingen for X skal vi først finde ud, hvilke værdier X antager for de enkelte udfald i udfaldsrummet. Det er gjort skematisk på figuren nedenfor til venstre. Situationen er set fra spillerens synspunkt, så et tab anføres som et negativt tal. Vi ser, at X kan antage 6 forskellige værdier: –4, 0, 1, 2, 3, og 4.

Rød terning

6	-4	4	3	2	1	0
5	-4	3	2	1	0	1
4	-4	2	1	0	1	2
3	-4	1	0	1	2	3
2	-4	0	1	2	3	4
1	-4	-4	-4	-4	-4	-4
	1	2	3	4	5	6

Grøn terning

Sandsynlighedsfordelingen for X:

x_i	–4	0	1	2	3	4
$P(X = x_i)$	$\frac{11}{36}$	$\frac{5}{36}$	$\frac{8}{36}$	$\frac{6}{36}$	$\frac{4}{36}$	$\frac{2}{36}$

Sandsynligheden for hver af disse værdier fås ved at tælle op, hvor ofte de forekommer i skemaet og så gange med 1/36, som er sandsynligheden for hvert enkelt udfald. Det giver tabellen til højre på forrige side.

□

2.4 Middelværdi, varians og spredning

Der er specielt to størrelser, som er med til at sige noget om en stokastisk variabel, og det er *middelværdien* og *variansen* eller *spredningen*. Middelværdien fås ved at tage det *vægtede gennemsnit* af værdierne for den stokastiske variabel, mens variansen fås som det vægtede gennemsnit af kvadratet på afvigelserne fra middelværdien. Spredningen er kvadratroden af variansen. Lad os regne på eksempel 2.21 og 2.22 ovenfor.

Eksempel 2.23

Middelværdien af den stokastiske variabel fra eksempel 2.21 fås ved at betragte tabellen for sandsynlighedsfordelingen og udregne det vægtede gennemsnit:

$$(5) \quad 2\cdot\tfrac{1}{36}+3\cdot\tfrac{2}{36}+4\cdot\tfrac{3}{36}+5\cdot\tfrac{4}{36}+6\cdot\tfrac{5}{36}+7\cdot\tfrac{6}{36}+8\cdot\tfrac{5}{36}+9\cdot\tfrac{4}{36}+10\cdot\tfrac{3}{36}+11\cdot\tfrac{2}{36}+12\cdot\tfrac{1}{36}$$

Denne udregning giver 7. Resultatet er ikke overraskende, da fordelingen er symmetrisk omkring 7. Middelværdien betegnes nogle gange med μ, andre gange med $E(X)$. Sidstnævnte står for "expectation of X" – den forventede værdi af X.

Nu til variansen af den stokastiske variabel X. Da $\mu = 7$, trækker vi 7 fra alle de værdier, som X kan antage, opløfter til anden potens og vægter med sandsynlighederne:

$$(6) \qquad \mathrm{Var}(X) \;=\; (2-7)^2\cdot\tfrac{1}{36}+(3-7)^2\cdot\tfrac{2}{36}+(4-7)^2\cdot\tfrac{3}{36}+\ldots+(12-7)^2\cdot\tfrac{1}{36} \;=\; 5{,}83$$

Spredningen fås ved at tage kvadratroden af variansen:

$$(7) \qquad \sigma \;=\; \sqrt{\mathrm{Var}(X)} \;=\; \sqrt{5{,}83} \;=\; 2{,}52$$

Selv om talværdien for spredningen ikke siger noget direkte om fordelingen, så kan man dog sige, at jo større tallet er, jo mere spredte er de værdier, som X antager.

□

Generelt defineres størrelserne således:

Definition 2.24 (Middelværdi, varians og spredning af stokastisk variabel)

Lad X være en stokastisk variabel på et endeligt sandsynlighedsfelt.
Med *middelværdien* eller *den forventede værdi* for X menes:

$$(8) \qquad \mu = E(X) = \sum_{i=1}^{n} x_i \cdot P(X = x_i)$$
$$= x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) + \ldots + x_n \cdot P(X = x_n)$$

Med *variansen* af X menes:

$$(9) \quad Var(X) = \sum_{i=1}^{n} (x_i - \mu)^2 \cdot P(X = x_i)$$
$$= (x_1 - \mu)^2 \cdot P(X = x_1) + (x_2 - \mu)^2 \cdot P(X = x_2) + \ldots + (x_n - \mu)^2 \cdot P(X = x_n)$$

Spredningen er kvadratroden af variansen:

$$(10) \qquad \sigma = \sigma(X) = \sqrt{Var(X)}$$

Der summeres over alle de værdier, som X kan antage: $x_1, x_2, \ldots, x_n$.

Eksempel 2.25

I spillet i eksempel 2.22 får vi følgende værdier for henholdsvis middelværdi og varians for den stokastiske variabel X:

$$\mu = E(X) = \sum_{i=1}^{n} x_i \cdot P(X = x_i)$$
$$= -4 \cdot \tfrac{11}{36} + 0 \cdot \tfrac{5}{36} + 1 \cdot \tfrac{8}{36} + 2 \cdot \tfrac{6}{36} + 3 \cdot \tfrac{4}{36} + 4 \cdot \tfrac{2}{36}$$
$$= -\tfrac{1}{9}$$

$$\mathrm{Var}(X) = \sum_{i=1}^{n} (x_i - \mu)^2 \cdot P(X = x_i)$$
$$= (-4 - (-\tfrac{1}{9}))^2 \cdot \tfrac{11}{36} + (0 - (-\tfrac{1}{9}))^2 \cdot \tfrac{5}{36} + (1 - (-\tfrac{1}{9}))^2 \cdot \tfrac{8}{36}$$
$$+ (2 - (-\tfrac{1}{9}))^2 \cdot \tfrac{6}{36} + (3 - (-\tfrac{1}{9}))^2 \cdot \tfrac{4}{36} + (4 - (-\tfrac{1}{9}))^2 \cdot \tfrac{2}{36}$$
$$= 7{,}65$$

Vi ser, at middelværdien er negativ, hvilket ikke er så mærkeligt, da bankører har det med at sørge for, at de selv har de bedste odds! Helt præcist fortæller middelværdien, at spilleren i gennemsnit vil *tabe* 1/9 kr. i hvert spil. Talværdien for variansen er ikke nem at give en god fortolkning af, men vi kan lave lidt om på spilreglerne og studere den virkning, som det har på variansen. Lad os sige, at spilleren stadig taber 4 kr., hvis blot en af terningerne viser 1, at to ens giver summen af øjnene i kr., undtagen hvis de to ens er (1,1), mens alle andre kombinationer hverken giver tab eller gevinst. Situationen er vist

på figuren på næste side. Middelværdien viser sig at være nøjagtig den samme som i det oprindelige spil, men variansen kan vises at være vokset til 14,87. Det skyldes, at spillet er blevet mere chancebetonet. Der er større præmier, som er fordelt på færre udfald. Men bankøren vil altså i gennemsnit få den samme indtjening!

Rød terning						
6	-4	0	0	0	0	12
5	-4	0	0	0	10	0
4	-4	0	0	8	0	0
3	-4	0	6	0	0	0
2	-4	4	0	0	0	0
1	-4	-4	-4	-4	-4	-4
	1	2	3	4	5	6

Grøn terning

□

Lige en definition, som dog især bruges, når den stokastiske variabel er normalfordelt eller binomialfordelt. Dem kigger vi på i næste kapitel:

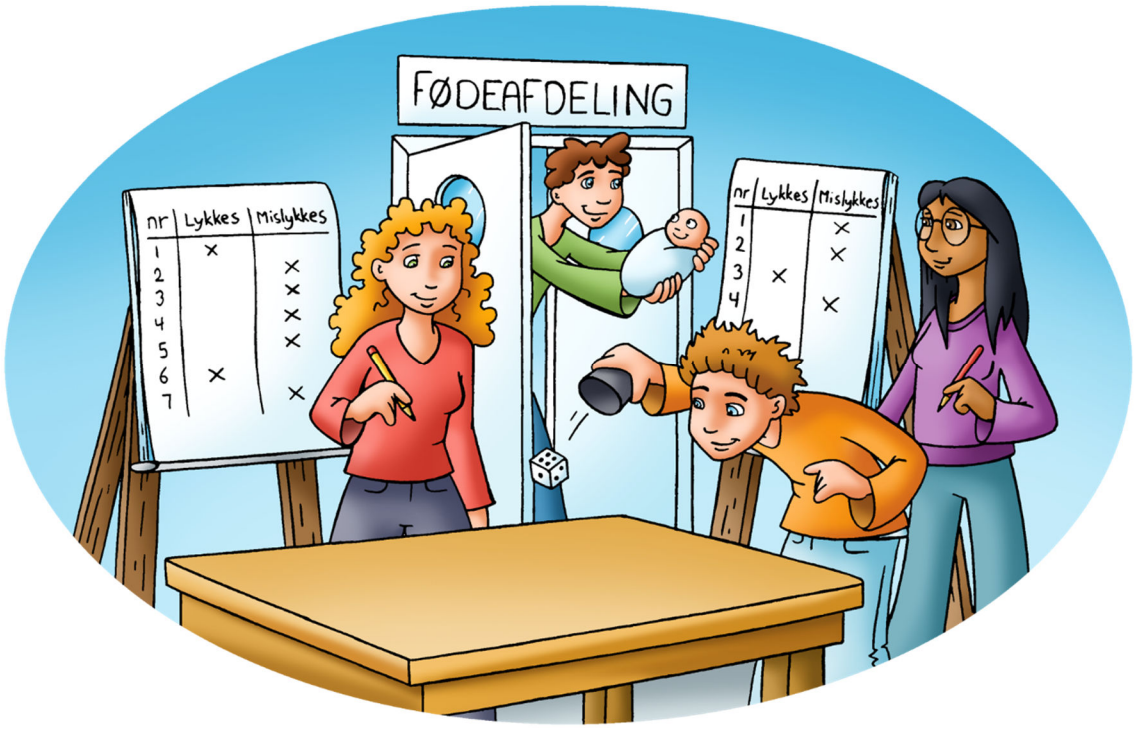
Definition 2.26 (Normale og exceptionelle værdier)

Lad X være en stokastisk variabel på et endeligt sandsynlighedsfelt.

En værdi x for den stokastiske variabel X kaldes *normal*, såfremt værdien højest ligger to spredninger fra middelværdien, altså hvis $\mu - 2\sigma \le x \le \mu + 2\sigma$.

En værdi x for den stokastiske variabel X kaldes *exceptionel*, såfremt værdien ligger mere end tre spredninger fra middelværdien, altså hvis $x < \mu - 3\sigma$ eller $x > \mu + 3\sigma$.

3. Binomialfordelingen



3.1 Nogle simple principper

Vi har allerede kigget på forskellige sandsynlighedsfordelinger i forrige kapitel. Disse fordelinger var imidlertid mere specielle i deres art. I dette kapitel skal vi studere en af de fordelinger, som finder allerstørst anvendelse i praksis, nemlig den såkaldte *binomialfordeling*. Årsagen til dens store udbredelse er, at den bygger på nogle meget generelle principper, som ofte er opfyldt i praktiske situationer:

Definition 3.1 (Binomialfordelingen)

Man har at gøre med et eksperiment, som består af n identiske *basiseksperimenter*. n kaldes for *antalsparameteren* eller eksperimentets *længde*. Udfaldene af de enkelte basiseksperimenter er indbyrdes *uafhængige*. Hvert basiseksperiment kan enten *lykkes* eller *mislykkes*. Sandsynligheden for, at basiseksperimentet lykkes, betegner vi med p, og p betegnes *basissandsynligheden* eller *sandsynlighedsparameteren*. Sandsynligheden for, at basiseksperimentet mislykkes, er dermed $1-p$. Den binomialfordelte stokastiske variabel X angiver, hvor mange gange basiseksperimentet lykkes.

Som det umiddelbart ses af definition 3.1, kan den binomialfordelte stokastiske variabel kun antage heltallige, ikke-negative værdier, nærmere bestemt tallene 0, 1, 2, … , n. Der er derfor tale om det, man kalder for en *diskret* stokastisk variabel. Et udfald u kan beskrives ved en vektor med n koordinater. Den i'te koordinat er lig L, hvis det i'te basiseksperiment lykkes og M, hvis det mislykkes. Eksempel: Hvis $n=5$, er $(\mathrm{L},\mathrm{M},\mathrm{M},\mathrm{L},\mathrm{M})$ det udfald, at første og fjerde basiseksperiment lykkes, mens de øvrige tre mislykkes. Mængden af alle udfald udgør udfaldsrummet. I det næste afsnit skal vi se, hvordan denne sandsynlighedsfordeling kommer til at se ud.

3.2 Binomialfordelingens sandsynligheder

Vi kan faktisk straks bestemme sandsynlighederne i binomialfordelingen, alene ud fra de abstrakte definitioner ovenfor. Når det er gjort, vil vi se på eksempler, som illustrerer fordelingens mangeartede anvendelse.

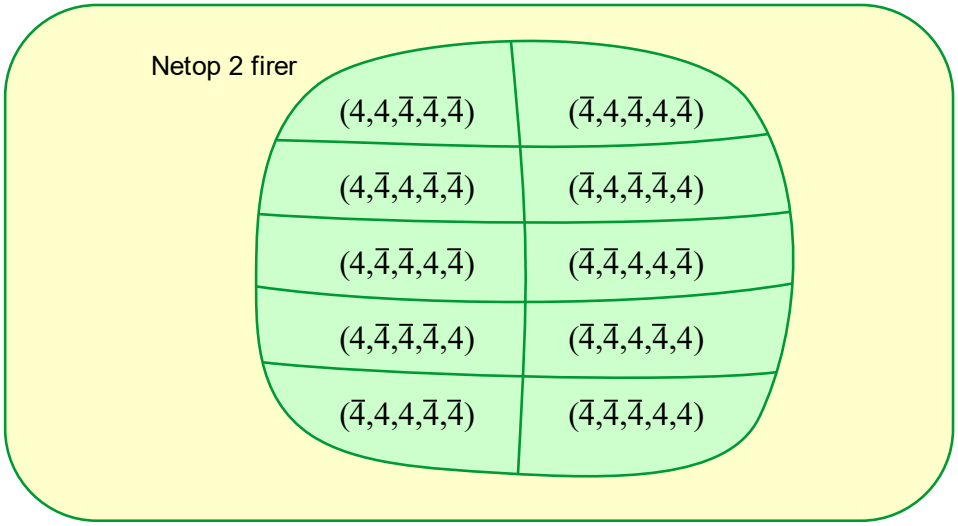
Sætning 3.2 (Binomialfordelingen)

Lad X være en binomialfordelt stokastisk variabel. Sandsynligheden for, at X antager værdien r, er givet ved følgende udtryk:

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

Bevisskitse: Af hensyn til overskueligheden vil vi argumentere for ovenstående formel ved hjælp af et eksempel: Lad os antage, at vi slår 5 gange med en terning og ønsker at finde sandsynligheden for at få netop 2 firer. Basiseksperimentet er da at slå én gang med

en terning. Vi vedtager, at basiseksperimentet lykkes, hvis man får en firer. Det giver en basissandsynlighed på $p=\frac{1}{6}$. Sandsynligheden for at basiseksperimentet mislykkes, dvs. at resultatet ikke er en firer, er dermed $1-p=\frac{5}{6}$. Den stokastiske variabel X angiver antallet af gange basiseksperimentet lykkes, dvs. hvor mange gange man får en firer i $n=5$ kast. At få netop to firer kan fremkomme på forskellig vis: Det er hensigtsmæssigt at dele op i en række hændelser. På figuren nedenfor symboliserer skrivemåden $(4,4,\overline{4},\overline{4},\overline{4})$, at man fik en firer i de to første kast, og *ikke-firer* i de næste tre kast. Man kan også forestille sig, at man fik en firer i første og tredje kast og ikke-firer i de øvrige tre kast, etc.



Der er i alt 10 muligheder, kan vi se, og de udelukker indbyrdes hinanden samtidigt med, at de udgør alle tilfælde med netop 2 firer. Med en matematisk betegnelse har vi at gøre med en såkaldt *klassedeling*, som gennemgået side 18 i kapitlet om kombinatorik. Sandsynligheden for at få netop to firer, kan da findes ved at lægge sandsynligheden for hver af de 10 muligheder sammen, jf. additionsprincippet. Lad os først prøve at bestemme sandsynligheden for $(4,4,\overline{4},\overline{4},\overline{4})$:

$$P(4,4,\overline{4},\overline{4},\overline{4}) = \tfrac{1}{6}\cdot\tfrac{1}{6}\cdot\tfrac{5}{6}\cdot\tfrac{5}{6}\cdot\tfrac{5}{6} = \left(\tfrac{1}{6}\right)^2\cdot\left(\tfrac{5}{6}\right)^3$$

Hvor vi har benyttet, at de enkelte kast er uafhængige af hinanden, så vi får sandsynligheden for hele kombinationen ved at gange sandsynlighederne for delhændelserne sammen. Vi kan gøre det samme med hændelsen $(4,\overline{4},4,\overline{4},\overline{4})$:

$$P(4,\overline{4},4,\overline{4},\overline{4}) = \tfrac{1}{6}\cdot\tfrac{5}{6}\cdot\tfrac{1}{6}\cdot\tfrac{5}{6}\cdot\tfrac{5}{6} = \left(\tfrac{1}{6}\right)^2\cdot\left(\tfrac{5}{6}\right)^3$$

Vi ser, at resultatet er det samme som før. Man overbeviser sig hurtigt om, at alle disse sandsynligheder er ens, nemlig $\left(\frac{1}{6}\right)^2\cdot\left(\frac{5}{6}\right)^3$. Derfor fås sandsynligheden for netop 2 firer ved at gange $\left(\frac{1}{6}\right)^2\cdot\left(\frac{5}{6}\right)^3$ med antallet af kombinationer:

$$P(X=2) = 10\cdot\left(\tfrac{1}{6}\right)^2\cdot\left(\tfrac{5}{6}\right)^3 = K(5,2)\cdot\left(\tfrac{1}{6}\right)^2\cdot\left(\tfrac{5}{6}\right)^{5-2}$$

idet $K(5,2)=10$ er antallet af kombinationer, svarende til antallet af måder, hvorpå man kan "udpege" de to pladser, hvor firerne skal stå. Resten af pladserne skal indeholde ikke-firer. Vi ser, at resultatet stemmer med (1) for $n=5,\ r=2,\ p=\frac{1}{6}$.

□

Eksempel 3.3 (Terningekast)

Bestem sandsynligheden for ved 8 kast med en terning at få netop 3 femmere. Angiv i den forbindelse desuden hele sandsynlighedsfordelingen for den stokastiske variabel X, som angiver antal femmere ved de 8 kast.

Løsning: Basiseksperimentet er ét kast med én terning, og det udføres $n = 8$ gange. Udfaldene af de enkelte basiseksperimenter er klart uafhængige. Dermed kan vi benytte binomialfordelingen. Basissandsynligheden er $p = \frac{1}{6}$. Ved brug af formel (1) fås:

$$P(X = 3) = K(8,3) \cdot \left(\tfrac{1}{6}\right)^3 \left(\tfrac{5}{6}\right)^5 = 56 \cdot \left(\tfrac{1}{6}\right)^3 \left(\tfrac{5}{6}\right)^5 = 0{,}1042$$

Så den søgte sandsynlighed er altså 10,4%. På helt tilsvarende måde kan man udregne sandsynligheden for alle de øvrige mulige værdier, som den stokastiske variabel kan antage, og derefter anbringe resultaterne i et skema. Hermed haves den såkaldte *sandsynlighedsfordeling* for den stokastiske variabel X:

r	0	1	2	3	4	5	6	7	8
$P(X = r)$	0,2326	0,3721	0,2605	0,1042	0,0261	0,0042	0,0004	0,0000	0,0000

Det er klart, at de forskellige CAS-værktøjer har kommandoer eller faciliteter til at udregne disse sandsynligheder hurtigt.

□

Eksempel 3.4 (Barnefødsler)

Det er en udbredt opfattelse blandt folk, at hvis man får mange børn af samme køn, så er det en ekstrem hændelse. Men er det nu det? Lad os se på eksempler. Hvad er sandsynligheden for ved 4 fødsler at få: a) fire piger? b) mindst en pige? c) højst to piger?

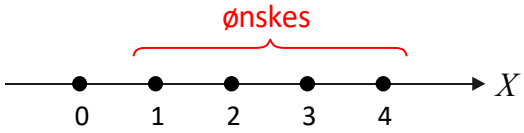
Løsning: Basiseksperimentet er én fødsel. Det oplyses, at det at få en pige eller en dreng kan betragtes som uafhængige hændelser. Hermed menes, at sandsynligheden for at få en dreng eller pige er helt uafhængigt at hvad kønnet på eventuelt tidligere børn måtte have været. Statistikker viser, at piger forekommer en smule sjældnere end drenge, nemlig i 49% af tilfældene. Lad os vedtage, at basiseksperimentet lykkes, hvis det bliver en pige, dvs. basissandsynligheden er $p = 0{,}49$. Lad X angive antallet af piger. Det er klart, at $n = 4$, da basiseksperimentet udføres 4 gange.

a) Af sætning 3.2 fås:

$$P(X=4)=K(4,4)\cdot 0{,}49^4\cdot(1-0{,}49)^{4-4}=K(4,4)\cdot 0{,}49^4\cdot 0{,}51^0=0{,}49^4=0{,}0576$$

så fire piger ved 4 fødsler sker altså trods alt i knap 6% af alle tilfælde!

b) Hvis der skal være mindst én pige, skal X altså være lig med 1, 2, 3 eller 4. Problemet kan dermed løses ved at udregne $P(X=1)+P(X=2)+P(X=3)+P(X=4)$.

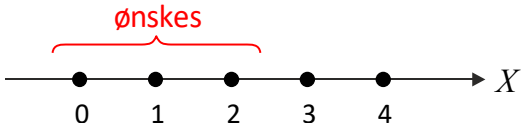


Der er imidlertid en nemmere måde: Det modsatte af at få mindst 1 pige er, at man *ingen* piger får (komplementære hændelse). Derfor kan opgaven også løses ved:

$$P(X\geq 1) \;=\; 1-P(X=0) \;=\; 1-K(4,0)\cdot 0{,}49^0\cdot 0{,}51^4 \;=\; 1-0{,}51^4 \;=\; 0{,}932$$

så sandsynligheden for at få mindst én pige ved 4 fødsler er altså 93,2%.

c) Hvis man skal have højst 2 piger, så skal X være lig med 0, 1 eller 2. Opgaven kan altså løses ved at udregne $P(X\leq 2)=P(X=0)+P(X=1)+P(X=2)$.



Bruger man formlen i sætning 3.2 på de tre sandsynligheder, fås:

$$P(X=0)=0{,}0677$$
$$P(X=1)=0{,}2600$$
$$P(X=2)=0{,}3747$$

Sammenlagt giver det en sandsynlighed på 70,2% for, at der ved fire fødsler højst forekommer to piger.

□

Bemærkning 3.5 (CAS-værktøj)

Det er klart, at det er regnetungt at foretage de beregninger, vi har gjort ovenfor. Det giver dog indsigt i matematikken, og er godt for forståelsen. Når man skal regne opgaver med større hjælpemidler, såsom et CAS-værktøj, er der heldigvis normalt altid et ekstra værktøj til rådighed i form af de *kumulerede binomial-sandsynligheder*. For overskuelighedens skyld kalder vi et sådan værktøj for *bincdf*:

$$(2)\quad bincdf(n,p,r) \;=\; P(X=0) \;+\; P(X=1) \;+\; \ldots \;+\; P(X=r) \;=\; \sum_{k=0}^{r} P(X=k)$$

hvor alle binomialsandsynlighederne fra 0 og op til r er lagt sammen. $bincdf(n,p,r)$, som funktion af den variable r, kaldes binomialfordelingens *fordelingsfunktion*.

Et direkte værktøj til at udregne punktsandsynligheder vil vi tilsvarende betegne *binpdf*:

(3) $$binpdf(n,p,r) = P(X=r)$$

Med disse to værktøjer bliver det meget hurtigere at løse opgaver i binomialfordelingen.

□

Sætning 3.6 (Middelværdi, varians og spredning)

Lad X være en binomialfordelt stokastisk variabel med antalsparameter n og sandsynlighedsparameter p. Ofte skrives det kort således: $X \sim b(n,p)$. *Middelværdien*, *variansen* og *spredningen* for X kan da udregnes ud fra følgende formler:

a) $\mu = E(X) = n \cdot p$

b) $Var(X) = n \cdot p \cdot (1-p)$

c) $\sigma = \sigma(X) = \sqrt{Var(X)} = \sqrt{n \cdot p \cdot (1-p)}$

Bevis: Beviser for ovenstående formler tager naturligvis udgangspunkt i den generelle definition 2.24. Da beviserne imidlertid er temmelig tekniske og ikke giver nogen særlig indsigt i øvrigt, så dropper vi dem.

□

At middelværdien er $n \cdot p$ er ikke så underligt. Hvis vi for eksempel slog 18 gange med en terning og var interesseret i antallet af femmere, så ville de fleste nok gætte på, at man i gennemsnit eller middel vil få $18 \cdot \frac{1}{6} = 3$ femmere. Det er netop hvad formlen også siger! Der er en sætning mere, som vi skal fremsætte uden bevis. Den handler om den mest sandsynlige værdi for X, svarende til den værdi for X, som har den højeste pind i et pindegram for sandsynlighedsfordelingen.

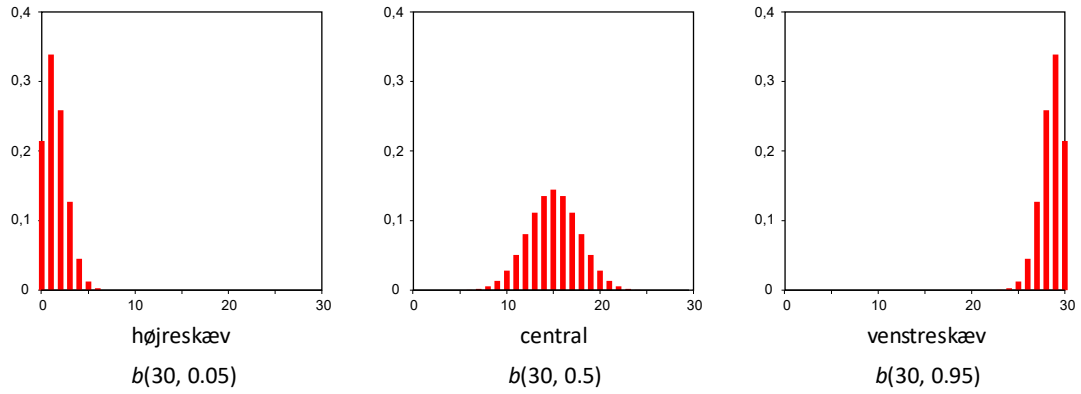
Sætning 3.7 (Mest sandsynlige udfald)

Givet en binomialfordelt stokastisk variabel X med antalsparameter n og sandsynlighedsparameter p. Hvis middelværdien $\mu = n \cdot p$ er et helt tal, så er middelværdien det mest sandsynlige udfald for X. Hvis μ *ikke* er et helt tal, så er det mest sandsynlige udfald af X ét af de to hele tal nærmest og på hver side af μ.

Definition 3.8 (Venstreskæv, højreskæv og central binomialfordeling)

En binomialfordeling kaldes *højreskæv*, hvis $\mu < 5$.
En binomialfordeling kaldes *venstreskæv*, hvis $\mu > n-5$.
I alle andre tilfælde kaldes binomialfordelingen *central*.

Nedenfor er afbildet nogle illustrationer af de forskellige typer binomialfordelinger i definition 3.8. På den første delfigur er sandsynlighedsparameteren lille, og der ses en hale mod højre. Den er *højreskæv*. Omvendt er sandsynlighedsparameteren for fordelingen vist på tredje delfigur stor og man ser en hale mod venstre. Den er *venstreskæv*. På den anden delfigur har vi en fordeling, som er mere symmetrisk. Den er *central*.



Det er på tide at se på et længere eksempel. Nu da vi har set brugen af formlen for binomialfordelingens punktsandsynligheder fra sætning 3.2, vil vi spare tid og bruge de værktøjer, som er omtalt i bemærkning 3.5, og som er til rådighed i de fleste CAS-værktøjer.

Eksempel 3.9

Ved folketingsvalget i 2019 var der 7,7% af danskerne, som stemte på partiet SF. Umiddelbart efter valget udspørges på tilfældigvis vis 50 danskere om, hvad de stemte på ved valget. Løs nedenstående spørgsmål i relation hertil.

a) Kan binomialfordelingen bruges til at løse opgaven?
b) Bestem sandsynligheden for, at netop 7 af de adspurgte stemte på SF.
c) Bestem sandsynligheden for at højst 4 stemte på SF.
d) Bestem sandsynligheden for at mindst 7 stemte på SF.
e) Bestem sandsynligheden for, at mindst 3 og højst 6 stemte på SF.
f) Hvor mange SF-stemmer vil der i gennemsnit være i en sådan stikprøve?
g) Bestem variansen og spredningen.
h) Hvad er det mest sandsynlige antal SF-stemmer?
i) Er binomialfordelingen venstreskæv, højreskæv eller central?
j) Afgør om værdierne 5 og 9 SF-stemmer er et normalt udfald, et exceptionelt udfald eller ingen af delene ifølge definition 2.26.

Løsninger:

a) Ja binomialfordelingen kan bruges. Basiseksperimentet består i at udspørge én person. Basiseksperimentet lykkes, hvis personen stemmer på SF og mislykkes, hvis denne ikke gør det. Den stokastiske variabel X angiver hvor mange gange basiseksperimentet lykkes, dvs. antallet af personer, som stemmer på SF. Antalsparameteren n er antallet af udspurgte og sandsynlighedsparameteren p er 7.7% eller 0.077. Be-

mærk, at i princippet er der ikke helt samme sandsynlighed fra det ene basiseksperiment til det andet, for hvad enten det oplyses, at den adspurgte person stemte på SF eller ej, så ændrer det en lille bitte smule på sandsynligheden for, at det næste basiseksperiment lykkes – fordi vi kun kender den samlede procentvise stemmeprocent på SF! Da vi imidlertid kun udspørger en ganske lille del af befolkningen, kan vi se bort fra dette. Vi kan med andre ord i praksis gå ud fra, at der er samme basissandsynlighed i hvert basiseksperiment!

b) Vi skal udregne en punktsandsynlighed og bruger derfor værktøjet *binpdf*.

$$binpdf(50, 0.077, 7) = 0{,}05112177042$$

Der er altså 5.1% sandsynlighed for, at der er netop 7 personer, som stemmer på SF.

c) Her skal vi have de kumulerede sandsynligheder i spil. Højst 4 stemmer på SF betyder, at vi skal udregne $P(X \le 4)$. Det er værktøjet *bincdf*, vi skal bruge:

$$bincdf(50, 0.077, 4) = 0{,}6594170427$$

Der er altså en sandsynlighed på 65.9% for, at der højst er 4 stemmer på SF i stikprøven med 50 stemmer.

d) Når man skal udregne sandsynligheder for mindst et givet tal, så er det mest fornuftigt at vende problemet på hovedet og bruge sætning 2.7 med den komplementære hændelse. Det modsatte af, at der mindst er 7 personer, som stemmer på SF, er, at der højst er 6 personer, som stemmer på SF.

$$1 - bincdf(50, 0.077, 6) = 0{,}0872765146$$

Der er altså 8.7% sandsynlighed for at mindst 7 personer stemmer på SF.

e) Vi udregner først sandsynligheden for, at der er højst 6 SF-stemmer og trækker derefter sandsynligheden for højst 2 SF-stemmer fra:

$$bincdf(50, 0.077, 6) - bincdf(50, 0.077, 2) = 0{,}6634380940$$

Der er altså 66.3% sandsynlighed for, at der er mindst 3 og højst 6 stemmer på SF.

f) Middelværdien af X:

$$\mu = E(X) = n \cdot p = 50 \cdot 0{,}077 = 3{,}85$$

Der vil altså i gennemsnit være 3.9 SF stemmer i en stikprøve på 50 personer.

g) Variansen beregnes ved hjælp af formlen i sætning 3.6:

$$Var(X) = n \cdot p \cdot (1-p) = 50 \cdot 0{,}077 \cdot (1-0{,}077) = 3{,}553550$$

Variansen er altså 3.55.

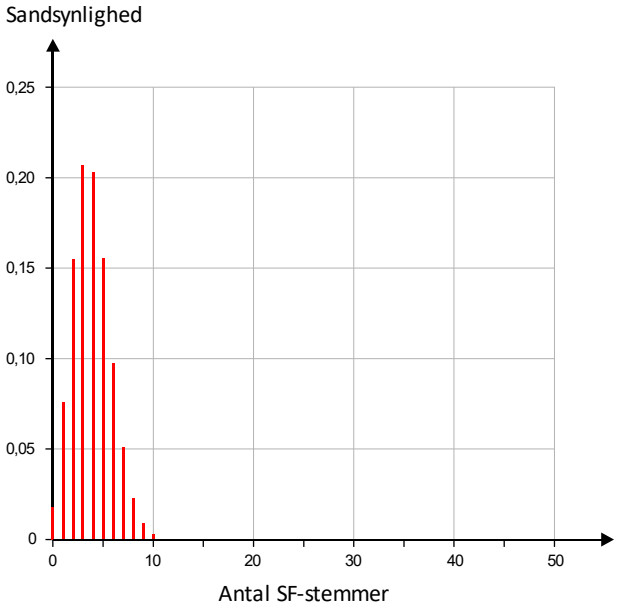
$$\sigma = \sqrt{Var(X)} = \sqrt{3{,}553550} = 1{,}885086205$$

Spredningen er altså 1.9 personer.

h) Da middelværdien af X ikke giver et helt tal, så må vi ifølge sætning 3.7 undersøge punktsandsynlighederne i de to hele værdier rundt omkring middelværdien på 3.85. Derfor undersøger vi 3 og 4:

$$bincdf(50, 0.077, 3) = 0{,}2071133985$$
$$bincdf(50, 0.077, 4) = 0{,}2030182528$$

Da punktsandsynligheden for 3 er størst, konkluderer vi, at det mest sandsynlige antal SF-stemmer er 3. Det kan også ses ved at lave et pindediagram. Diagrammet er vist på næste side, og vi ser tydeligt, at pinden ud for 3 SF-stemmer er højest.



i) Eftersom $\mu = 3{,}85 < 5$, er binomialfordelingen for X *højreskæv*!

j) Normalområdet er bestemt af intervallet imellem følgende værdier:

$$\mu - 2\sigma = 0{,}079827590$$
$$\mu + 2\sigma = 7{,}620172410$$

Da værdien 5 ligger i intervallet mellem disse to tal, konkluderer vi, at 5 SF-stemmer er en normal observation.

Det exceptionelle område ligger udenfor intervallet bestemt af disse værdier:

$$\mu - 3\sigma = -1{,}805258615$$
$$\mu + 3\sigma = 9{,}505258615$$

Da 9 ligger *indenfor* (ikke udenfor) intervallet bestemt af ovenstående værdier, konkluderer vi, at 9 SF stemmer hverken er en normal eller exceptionel observation.

□

Bemærkning 3.10 (Med eller uden tilbagelægning)

Der er en kommentar, som er på sin plads her, og det er problematikken omkring *med tilbagelægning* og *uden tilbagelægning*. Som bekendt kræver binomialfordelingen, at der er samme basissandsynlighed, hver gang et nyt basiseksperiment udføres. Det er intet problem ved terningekast, men hvis det handlede om at trække et kort fra et kortspil, og man ikke lægger kortet tilbage, så ville det være et problem, da sandsynlighederne så vil ændre sig. Så skal helt andre fordelinger benyttes. I nogle tilfælde kan man dog se bort fra den begåede fejl, fx som da vi i eksempel 3.9 trak fra en meget stor pulje!

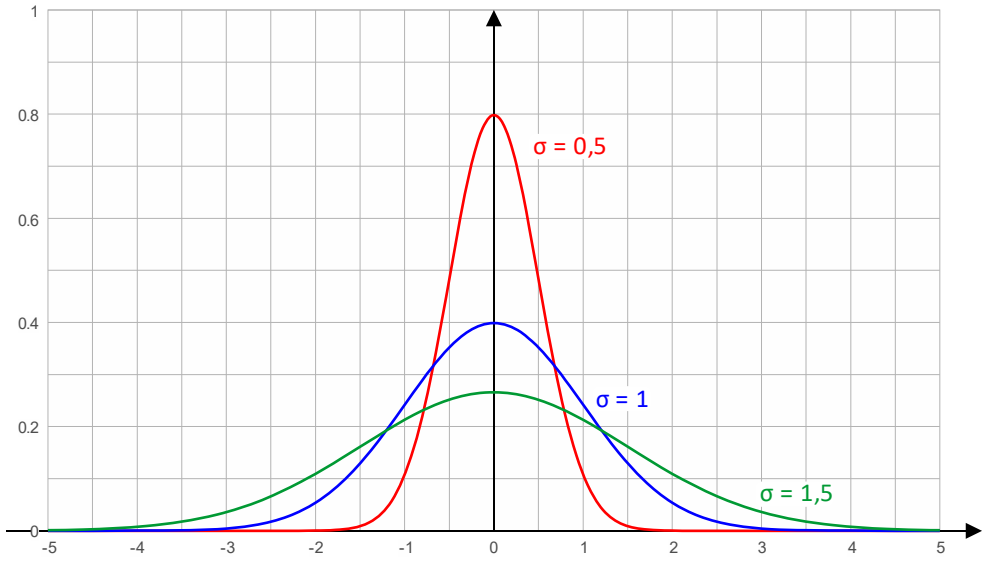
3.3 Approksimation med normalfordelingen

Normalfordelingen er den mest anvendte sandsynlighedsfordeling. Der er tale om en såkaldt *kontinuert* fordeling, forstået på den måde, at en normalfordelt stokastisk variabel kan antage værdier i et helt interval. Den er defineret via en *frekvensfunktion* eller *tæthedsfunktion* givet ved

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\cdot\sqrt{2\pi}}\cdot e^{-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2} \tag{4}$$

Vi ser, at den har to parametre, nemlig μ og σ, som viser sig at repræsentere middelværdien og spredningen for fordelingen. På figuren nedenfor er afbildet grafen for tæthedsfunktionen for forskellige værdier af σ. Vi ser, at jo større σ er, jo bredere er *klokkekurven*, som den ofte omtales. Alle kurver nedenfor har $\mu = 0$. Der er ingen grund til at vise variationen i middelværdien μ, for den eneste virkning den parameter har er, at den parallelforskyder grafen langs x-aksen, så klokkekurven er symmetrisk omkring μ.

Graferne for tre tæthedsfunktioner for normalfordelingen (alle μ = 0)



$X \sim N(\mu,\sigma)$ hentyder til, at man har at gøre med en normalfordelt stokastisk variabel med parametre μ og σ. Det giver ikke mening at bede om punktsandsynligheder, når der som her er tale om en kontinuert fordeling. Derimod kan man udregne sandsynligheden for at den normalfordelte stokastiske variabel X antager en værdi i et interval. Sandsynligheden udregnes ved at integrere tæthedsfunktionen over det pågældende interval:

$$P(a \le X \le b) = \int_a^b f_{\mu,\sigma}(z)dz = \int_a^b \frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot e^{-\frac{1}{2}\cdot\left(\frac{z-\mu}{\sigma}\right)^2} dz \tag{5}$$

Integrerer man fra $-\infty$ til ∞ får man 1, som forventet. Specielt har vi fordelingsfunktionen, som angiver sandsynligheden for at X er mindre end eller lig med en given værdi x.

$$F_{\mu,\sigma}(x) = P(X \le x) = \int_{-\infty}^{x} f_{\mu,\sigma}(z)dz = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot e^{-\frac{1}{2}\cdot\left(\frac{z-\mu}{\sigma}\right)^2} dz \tag{6}$$

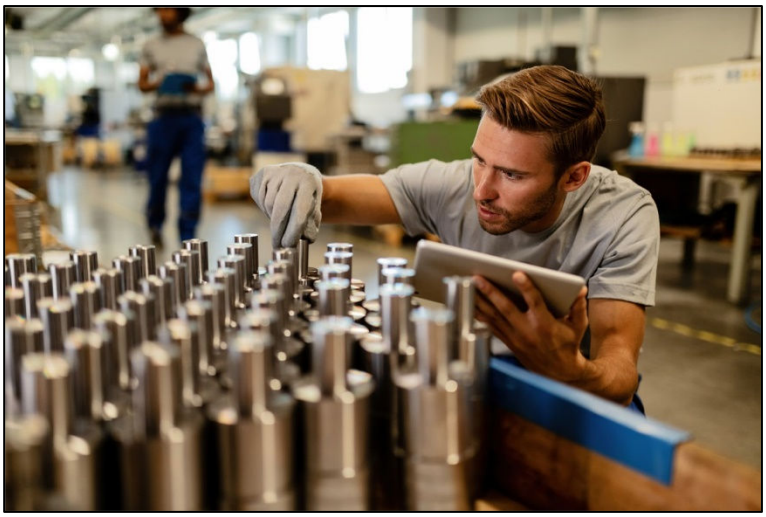
Af (5) og (6) fås umiddelbart:

$$P(a \le X \le b) \;=\; F_{\mu,\sigma}(b) - F_{\mu,\sigma}(a) \tag{7}$$

som er en stor fordel i mange sammenhænge, da man så ikke behøver integrere.

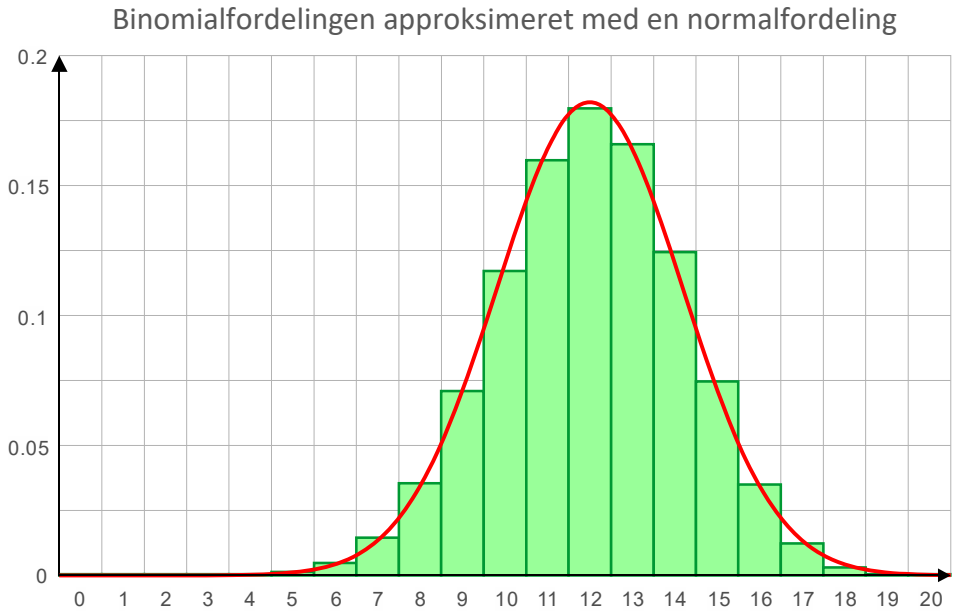
Læseren kan måske undre sig over, hvorfor tæthedsfunktionen ser ud som den gør og hvorfor fordelingen spiller så stor en rolle i sandsynlighedsregning. Der er ingen nem forklaring på det. Fordelingen blev til via nogle af de største matematikere gennem tiderne, herunder Carl Friedrich Gauss (1777-1855) og Pierre-Simon Laplace (1749-1823), og det skete ad kringlede veje. Normalfordelingen blev på et tidspunkt set som en approksimation til binomialfordelingen. Normalfordelingskurven blev også betragtet som en *fejlkurve* til at beskrive fordelingen af usikkerheder ved målinger af en fysisk størrelse.

Der er en lang række eksempler fra det virkelige liv, hvor man har erfaringer for at data tilnærmelsesvist er normalfordelte. Et godt eksempel er højden af soldater målt ved session. Ved produktion af komponenter i industrien, falder komponenterne ikke altid helt ens ud i størrelse eller vægt. Når flere små tilfældige og indbyrdes uafhængige effekter er til stede i en produktionsproces, så har man praktisk erfaring for, at produkterne tilnærmelsesvist følger en normalfordeling på et eller flere punkter. Dette er også underbygget teoretisk gennem den såkaldte *centrale grænseværdisætning*, som blev bevist af førnævnte Pierre-Simon Laplace i 1810 i en første version. Der er tale om en dyb sætning, der står som en hjørnesten i sandsynlighedsregningen.



Det er ikke meningen, at vi skal studere normalfordelingen nøjere her på B-niveau STX. Derimod bliver emnet behandlet nærmere på A-niveau. Vi skal mest forklare hvorfor normalfordelingen kan bruges som en approksimation til binomialfordelingen. Binomialfordelingen er som bekendt en diskret fordeling, fordi en binomialfordelt stokastisk variabel kun kan antage endeligt mange værdier, nemlig 0, 1, … , n. Det er måske derfor overraskende, at den kan tilnærmes med normalfordelingen, som jo er en kontinuert fordeling. Som et eksempel kan vi se på en binomialfordelt stokastisk variabel X med antalsparameter $n = 20$ og basissandsynlighed $p = 0{,}6$. På figuren på næste side er sandsynlighedsfordelingen for X afbildet, så søjlerne har centrum i de værdier, de hører til. Fra teorien

om binomialfordelingen ved vi, at der for en binomialfordelt stokastisk variabel X gælder, at middelværdien er givet ved følgende: $E(X) = n \cdot p = 20 \cdot 0{,}6 = 12$, mens variansen er givet ved følgende: $\mathrm{Var}(X) = n \cdot p \cdot (1-p) = 4{,}8$. På nævnte figur er desuden indtegnet tæthedsfunktionen for normalfordelingen med de samme værdier for middelværdi og varians, altså henholdsvis 12 og 4,8. Vi ser, at approksimationen er overraskende god!



Som en tommelfingeregel er normalfordelingen en rimelig god approksimation til binomialfordelingen, hvis følgende kriterier er opfyldt: $n > 9 \cdot p/(1-p)$ samt $n > 9 \cdot (1-p)/p$. En begrundelse for kriteriet springer vi over. Kort fortalt skal det dog nævnes, at den geniale matematiker *Abraham de Moivre* (1667-1754) søgte en måde at simplificere udregningerne af binomialsandsynligheder på. Husk på, at alt måtte regnes i hånden dengang! På den tid var normalfordelingen endnu ikke opdaget, men de resultater de Moivre kom frem til, kan tolkes på den måde, at han tilnærmede binomialfordelingen med en normalfordeling. Han kan samtidigt siges at være den person, som gav den første formulering af et specialtilfælde af den centrale grænseværdisætning. Du kan se en figur fra hans bog *Doctrine of Chances* på næste side.

Lad os til slut udtrykke approksimationen mere præcist. Den normalfordeling, som har middelværdi $\mu = 0$ og spredning $\sigma = 1$ kaldes for *standardnormalfordelingen*. Dens fordelingsfunktion kaldes $\Phi(x)$. Har man en tabel for den, kan man lynhurtigt få fordelingsfunktionen for enhver anden normalfordeling med parametre μ og σ, nemlig ved:

(8) $$F_{\mu,\sigma}(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

Der er tale om en substitution, men vi udelader detaljerne. Antag nu, at vi har givet en *binomialfordelt* stokastisk variabel $X \sim b(n,p)$. Den har som bekendt middelværdi givet ved $\mu = n \cdot p$ og spredning givet ved $\sigma = \sqrt{n \cdot p \cdot (1-p)}$. Da haves:

(9) $$P(X \le x) \approx F_{\mu,\sigma}\left(x+\tfrac{1}{2}\right) = \Phi\left(\frac{x+\frac{1}{2}-\mu}{\sigma}\right)$$

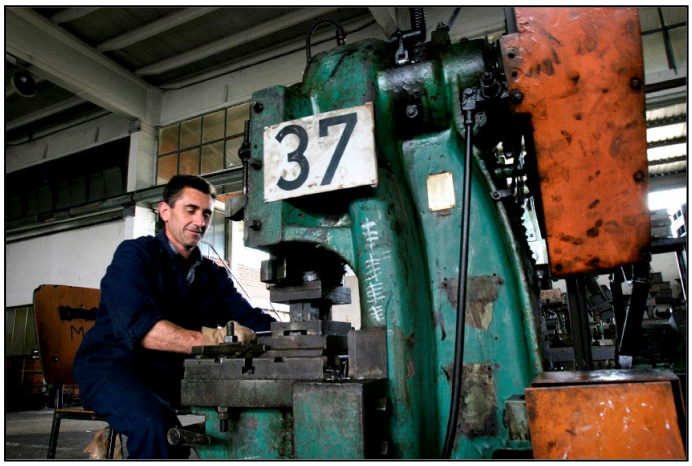
Grafen ovenfor antyder også en forklaring på, hvorfor der adderes $\frac{1}{2}$ (kontinuitetskorrektionen). Søjlerne er nemlig centreret midt i de talværdier, de repræsenterer. Så for at få den bedste approksimation, integreres fra $-\infty$ til $x+\frac{1}{2}$ snarere end fra $-\infty$ til x. Jo større n bliver, jo bedre er approksimationen. Fjerner man $\frac{1}{2}$, vil approksimationen dog også være fin, blot n ikke er for lille. Prøv eventuelt det hele af i et CAS-værktøj!

□

Som et spor til A-niveau gives der nu et lille eksempel på, hvordan normalfordelingen kan benyttes i en konkret og realistisk sammenhæng. Som nævnt tidligere er produktion af maskindele i industrien et godt eksempel på, hvor normalfordelingen kommer i spil.

Eksempel 3.11 (Variation i produktionen)

En maskine på en fabrik skal fremstille cylindre med en diameter på 20 mm. Imidlertid falder resultatet ikke altid helt nøjagtigt ud. Det viser sig, at diametrene er normalfordelte med middelværdi 20 mm og med en spredning på 0,1 mm. Fabrikanten kan acceptere en afvigelse på maksimalt 0,2 mm fra det ønskede.



a) Bestem sandsynligheden for, at en cylinder har en diameter på højst 19,8 mm.
b) Bestem sandsynligheden for en cylinderdiameter mellem 19,9 og 20,25 mm.
c) Hvor stor en del af cylindrene må kasseres?

Løsning:

a) For det første lader vi X betegne den normalfordelte stokastiske variabel, som angiver en cylinders diameter. Den har middelværdi $\mu = 20$ mm og spredning $\sigma = 0{,}1$ mm. Vi skal bestemme $P(X \le 19{,}8)$ og bruger (6):

$$P(X \le 19{,}8) = F_{20,0.1}(19{,}8) = 0{,}02275$$

Altså en sandsynlighed på ca. 2,3% for, at en cylinder har diameter højst 19,8 mm. Bemærk, at CAS-værktøjet typisk har fordelingsfunktionen indbygget. Det kan fx være, at den hedder *normalcdf* (cdf for den kumulerede sandsynlighed).

b) Vi får her:

$$P(19{,}9 \le X \le 20{,}25)$$
$$= P(X \le 20{,}25) - P(X \le 19{,}9) = F_{20,0.1}(20{,}25) - F_{20,0.1}(19{,}9) = 0{,}835135$$

… en sandsynlighed på 83,5% for, at en cylinderdiameter mellem 19,9 og 20,25 mm.

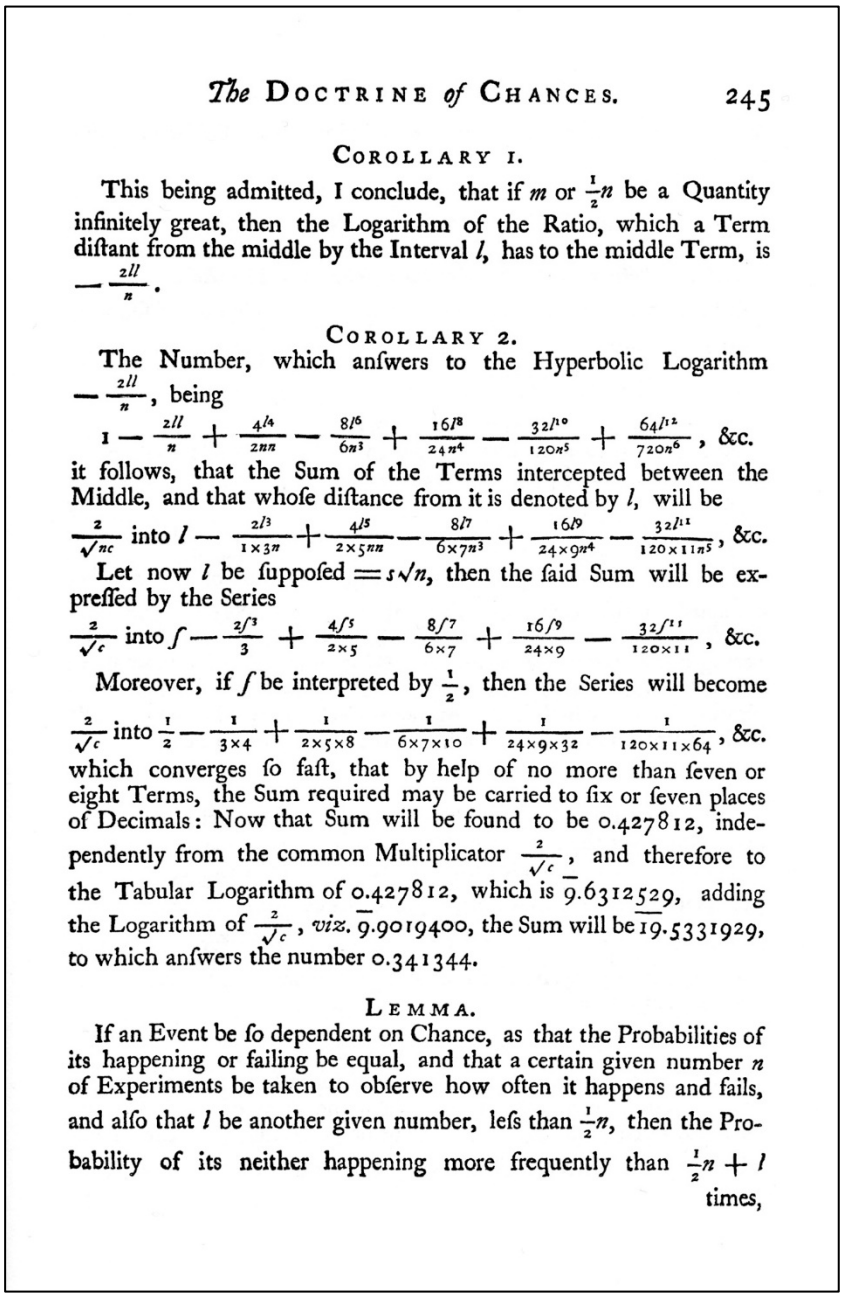
Bemærk, at vi med normalfordelingen ikke skal bekymre os om, hvorvidt en konkret værdi i et intervalendepunkt, her 19,9 mm, er med eller ej. Vi skal heller ikke længere til at trække 1 fra i det venstre intervalendepunkt, som tilfældet var med binomialfordelingen. Det skyldes, at normalfordelingen er en kontinuert fordeling!

c) En cylinder kasseres, hvis den enten har en diameter under 19,8 mm eller over 20,2 mm. Vi får følgende, idet vi blandt andet kigger på den komplementære hændelse.

$$P(X<19{,}8)+P(X>20{,}2) \;=\; P(X<19{,}8)+1-P(X\le 20{,}2)$$
$$=\; F_{20,0.1}(19{,}8)+1-F_{20,0.1}(20{,}2) \;=\; 0{,}045500$$

Altså en sandsynlighed på ca. 4,6% for, at en given cylinder skal kasseres.

□



The DOCTRINE *of* CHANCES. 245

COROLLARY I.

This being admitted, I conclude, that if m or $\frac{1}{2}n$ be a Quantity infinitely great, then the Logarithm of the Ratio, which a Term diſtant from the middle by the Interval l, has to the middle Term, is $-\frac{2ll}{n}$.

COROLLARY 2.

The Number, which anſwers to the Hyperbolic Logarithm $-\frac{2ll}{n}$, being

$$1-\frac{2ll}{n}+\frac{4l^4}{2nn}-\frac{8l^6}{6n^3}+\frac{16l^8}{24n^4}-\frac{32l^{10}}{120n^5}+\frac{64l^{12}}{720n^6}, \text{ \&c.}$$

it follows, that the Sum of the Terms intercepted between the Middle, and that whoſe diſtance from it is denoted by l, will be

$$\frac{2}{\sqrt{nc}} \text{ into } l-\frac{2l^3}{1\times 3n}+\frac{4l^5}{2\times 5nn}-\frac{8l^7}{6\times 7n^3}+\frac{16l^9}{24\times 9n^4}-\frac{32l^{11}}{120\times 11n^5}, \text{ \&c.}$$

Let now l be ſuppoſed $= s\sqrt{n}$, then the ſaid Sum will be expreſſed by the Series

$$\frac{2}{\sqrt{c}} \text{ into } s-\frac{2s^3}{3}+\frac{4s^5}{2\times 5}-\frac{8s^7}{6\times 7}+\frac{16s^9}{24\times 9}-\frac{32s^{11}}{120\times 11}, \text{ \&c.}$$

Moreover, if s be interpreted by $\frac{1}{2}$, then the Series will become

$$\frac{2}{\sqrt{c}} \text{ into } \frac{1}{2}-\frac{1}{3\times 4}+\frac{1}{2\times 5\times 8}-\frac{1}{6\times 7\times 10}+\frac{1}{24\times 9\times 32}-\frac{1}{120\times 11\times 64}, \text{ \&c.}$$

which converges ſo faſt, that by help of no more than ſeven or eight Terms, the Sum required may be carried to ſix or ſeven places of Decimals: Now that Sum will be found to be 0.427812, independently from the common Multiplicator $\frac{2}{\sqrt{c}}$, and therefore to the Tabular Logarithm of 0.427812, which is $\overline{9}.6312529$, adding the Logarithm of $\frac{2}{\sqrt{c}}$, *viz.* $\overline{9}.9019400$, the Sum will be $\overline{19}.5331929$, to which anſwers the number 0.341344.

LEMMA.

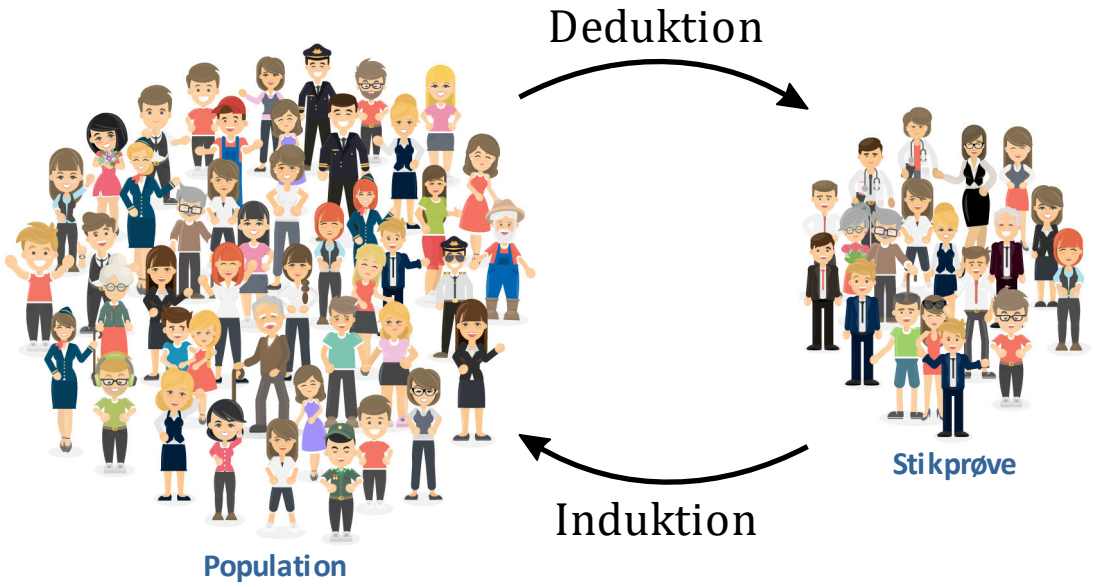
If an Event be ſo dependent on Chance, as that the Probabilities of its happening or failing be equal, and that a certain given number n of Experiments be taken to obſerve how often it happens and fails, and alſo that l be another given number, leſs than $\frac{1}{2}n$, then the Probability of its neither happening more frequently than $\frac{1}{2}n+l$ times,

Side 245 i Abraham De Moivres værk *The Doctrine of Chances*, 3. udgave 1756. Corollary 2 indeholder hovedresultatet.

4. Binomialtest og konfidensintervaller

4.1 Statistik overfor sandsynlighedsregning

I dette kapitel bevæger vi os over i statistikken, efter vi tidligere har arbejdet indenfor *kombinatorikken* og *sandsynlighedsregningen*. Mens kombinatorikken handler om at bestemme antallet af muligheder eller kombinationer, så beskæftiger sandsynlighedsregningen sig især med at udregne sandsynligheder, givet nogle bestemte forudsætninger. Det sidste kommer for eksempel til udtryk ved, at man udleder sandsynligheder for udfald i de forskellige *sandsynlighedsfordelinger*. Man kan tale om en *deduktiv* proces! *Statistik* er derimod mere en *induktiv* proces derved, at man typisk på baggrund af data fra den virkelige verden i form af en stikprøve forsøger at slutte sig til, hvilke fordelinger, som er relevante i den konkrete situation og/eller at bestemme eventuelle ukendte parametre i de anvendte sandsynlighedsfordelinger/modeller.



Det vil selvfølgelig være mest fordelagtigt, hvis man har fuld information om hele populationen. Det har Danmarks Statistik for eksempel på mange områder, da virksomheder og det offentlige har indberetningspligt. For andre analyseinstitutter og meningsmålingsinstitutter er det imidlertid alt for bekosteligt, ofte direkte umuligt at foretage undersøgelser på hele populationen. Her tages i stedet *stikprøver*, som man så håber er tilstrækkelig repræsentative for hele populationen. I statistikken behandles den usikkerhed, der følger med, når man begrænser sig til en stikprøve. Lidt løst sagt forsøger man at konkludere fra det specielle til det generelle! I denne proces er man dog nødt til også at benytte sig af sandsynlighedsteorien. Med den kan man nemlig udtale sig om den sandsynlighed, hvormed stikprøven er fremkommet! Modellen indikeret ved figuren ovenfor skal forstås bredt. Der behøver således ikke være tale om en population af personer, hvoraf der udtrækkes en delmængde af personerne. Populationen kan lige så godt være mængden af alle de produkter, som en virksomhed producerer, og stikprøven kan være et lille tilfældigt udvalg af disse. Populationen kan også være mængden af alle aktier, hvoraf en lille stikprøve udtages til analyse. I alle tilfælde ønsker man at udtale sig om hele populationen på baggrund af stikprøven.

4.2 Binomialtest

Vi vil begynde med at studere *hypotesetest*, specielt den type, som kaldes *binomialtest*. Det handler om, at man opstiller en nulhypotese, som gælder hele populationen. På baggrund af en stikprøve giver man derefter en vurdering af, om nulhypotesen mon er sand eller falsk. Enten accepterer man nulhypotesen eller også afviser man den. Før vi begynder at kommentere begrebet yderligere, er det hensigtsmæssigt at kigge på et eksempel, så vi har noget at hænge de generaliserende begreber op på. Og hvad er mere let at forstå end kast med en terning?

Eksempel 4.1 (Terningekast – tosidet)

Lad os sige, at vi er usikre på, om en given terning er uægte og ikke slår det antal seksere, man kan forvente, altså at terningen i det lange løb ikke viser en sekser i 1/6 af kastene. Vi gennemfører nu en stikprøve, hvor vi kaster 48 gange med en terning og får 3 seksere. Får det os til at tvivle på terningens ægthed? Hvis terningen er ægte, vil vi i gennemsnit i 48 kast forvente at få $48 \cdot \frac{1}{6} = 8$ seksere. Stikprøven viste med andre ord noget færre seksere end forventet. Vi har ikke andet data end stikprøven, så vi kan kun vurdere terningens ægthed ud fra udfaldet af stikprøven. Det er klart, at jo mere usædvanlig stikprøvens resultat er i forhold til det forventede, jo mere vil vi hælde til at sige, at terningen nok er uægte. Men hvad er "usædvanlig", og hvordan kvantificerer vi det? Vi skal stille det helt skarpt op ved at opstille en såkaldt *nulhypotese* H_0 samt en *alternativ* hypotese H_1:

H_0 : Terningen er ægte, dvs. den vil i det lange løb vise en sekser i 1/6 af kastene.
H_1 : Terningen er uægte, dvs. frekvensen af seksere er $\neq \frac{1}{6}$.

I 48 kast med en terning er det muligt at få netop 0 seksere, 1 sekser, 2 seksere, etc. op til maksimalt 48 seksere. Det er klart, at hvis nulhypotesen er sand, så vil det være ret usædvanligt at få meget få seksere eller rigtig mange seksere. Det er "halerne" i hver ende, som kommer til at udgøre vores *kritiske område*, og hvis det antal seksere, som vi slog, ligger i det kritiske område, vil vi afvise nulhypotesen. I modsat fald vil vi acceptere den. Spørgsmålet er hvor store disse haler i hver ende skal være. Her er idéen at lægge sig fast på et såkaldt *signifikansniveau* α. Det sættes ofte til 5% eller 0,05. Vi skærer nu signifikansniveauet over i to lige store dele og tildeler α/2 til hver hale. Lad os sige, at vi har valgt et signifikansniveau på 5%. Dermed vil der være 2,5% til hver hale. Det skal forstås på følgende måde: Hvad angår halen til venstre, så summerer vi binomialsandsynlighederne op fra bunden så længe den kumulerede sandsynlighed ikke overstiger 2,5%. De udfald, som er i spil, ligger i det kritiske område til venstre:

$$P(X=0) \; + \; P(X=1) \; + \; P(X=2) \; = \; 0{,}008817$$

hvor X naturligvis er den stokastiske variabel, som angiver antallet af seksere. Hvis vi havde lagt $P(X=3)$ til også, ville vi have fået summen 0,030711, som er over 0,025. Vi konkluderer, at den venstre hale leverer elementerne 0, 1 og 2 til det kritiske område. På tilsvarende vis har vi for den højre hale:

$$P(X=14) \; + \; P(X=15) \; + \; \ldots \; + \; P(X=48) \; = \; 0{,}021828$$

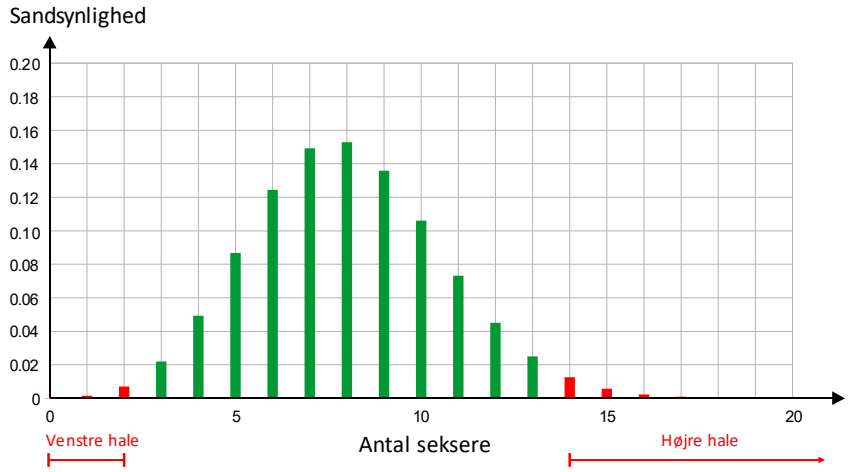
Havde vi taget $P(X=13)$ med, ville summen have været 0,046836, som er over 2,5%. Den højre hale leverer altså elementerne 14, 15, … , 48 til det kritiske område, som dermed kommer til at se således ud:

$$\text{Kritiske område} = \{0, 1, 2, 14, 15, \ldots, 48\}$$

hvilket automatisk giver anledning til *acceptområdet*:

$$\text{Acceptområdet} = \{3, 4, \ldots, 13\}$$

Da vores stikprøve indeholdt 3 seksere, og da 3 er indeholdt i acceptområdet, vælger vi at acceptere nulhypotesen ved et signifikansniveau på 5%. Den observerede værdi af X, som her altså er 3, benævnes *teststørrelsen*. Processen kan illustreres ved et pindediagram for sandsynlighedsfordelingen for den stokastiske variabel X. Pindenes højde angiver sandsynligheden for de respektive udfald. Pindene for udfaldene i det kritiske område er tegnet røde, mens pindene hørende til udfaldene i acceptområdet er tegnet grønne. Den samlede sandsynlighed for hver hale er maksimalt 2,5%.



□

Bemærkning 4.2

I et dobbeltsidet binomialtest er det venstre endepunkt k_v i acceptområdet lig med det *mindste* tal r, hvis kumulerede frekvens er *skarpt større* end $\alpha/2$: $bincdf(n, p, r) > \frac{1}{2}\alpha$. Her har vi anvendt betegnelsen *bincdf* for den kumulerede sandsynlighedsfunktion, som også omtalt i bemærkning 3.5 i forrige kapitel.

Det højre endepunkt k_h i acceptmængden for det dobbeltsidede binomialtest får på følgende måde: Lad $r_{\max}$ være den *største* værdi af r, for hvilket den kumulerede sandsynlighed er *skarpt mindre* end $1-\alpha/2$, altså $bincdf(n, p, r) < 1-\frac{1}{2}\alpha$. Da er $k_h = r_{\max} + 1$.

Det samlede acceptområde er dermed $\{k_v, \ldots, k_h\}$.

Det skal nævnes, at mange CAS-værktøjer har et værktøj, som direkte angiver acceptområdet $\{k_v, \ldots, k_h\}$, så man ikke skal bøvle med det.

□

Eksempel 4.3 (Terningekast – venstresidet)

Efter en god dags spil har en spiller fået en mistanke om, at en terning giver *for få* seksere. I det tilfælde kan det være fornuftigt at foretage et såkaldt *venstresidet* binomialtest. Nulhypotesen og den alternative hypotese, vil da se således ud:

H_0 : Terningen er ægte, dvs. den vil i det lange løb vise en sekser i 1/6 af kastene.

H_1 : Terningen giver for få seksere, dvs. frekvensen af seksere er $< \frac{1}{6}$.

Lad os sige, at udfaldet af vores stikprøve er nøjagtigt magen til den i eksempel 4.1, dvs. at der blev kastet 48 gange med terningen, og at teststørrelsen igen var 3 seksere. Det giver i dette tilfælde kun mening at tale om kritiske værdier til venstre. Mange seksere vil umiddelbart ikke få os til at tvivle på, om nulhypotesen er rigtig. Med et signifikansniveau på 5% tildeler vi alle 5% til den venstre hale! Vi søger den mindste værdi for *r*, for hvilket den kumulerede frekvens er skarpt større end $\alpha = 0{,}05$. Vi har:

$$P(X=0) + P(X=1) + P(X=2) + P(X=3) = 0{,}030711$$

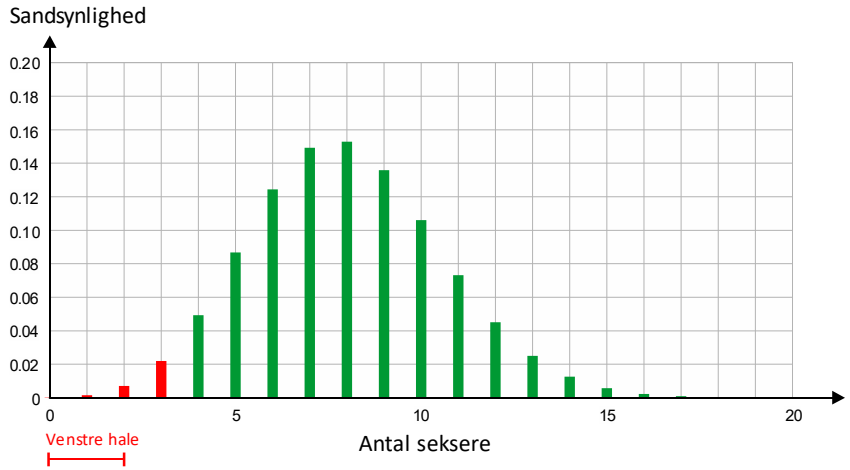
Havde vi medtaget $P(X=4)$, ville den kumulerede sandsynlighed have været 0,079974, som er over 0,05. Vi konkluderer, at det kritiske område udgøres af udfaldene 0, 1, 2 og 3. Det giver alt i alt:

$$\text{Kritiske område} = \{0, 1, 2, 3\}$$

mens acceptområdet bliver:

$$\text{Acceptområdet} = \{4, 5, \ldots, 48\}$$

Da teststørrelsen er 3 seksere, og da 3 er indeholdt i det kritiske område, vælger vi at afvise nulhypotesen ved et signifikansniveau på 5%. Vi ser hermed, at det godt kan forekomme, at man accepterer en nulhypotese i en tosidet test, mens man med den samme stikprøve afviser nulhypotesen, hvis der er valgt en venstresidet test. Nedenstående pindediagram viser situationen med de kritiske værdier i den venstre hale:



□

Bemærkning 4.4

I et venstresidet binomialtest fås det venstre endepunkt k_v i acceptområdet lig med det *mindste* tal *r*, hvis kumulerede frekvens er *skarpt større* end α, altså $bincdf(n,p,r) > \alpha$. Acceptområdet er dermed $\{k_v, \ldots, n\}$.

Eksempel 4.5 (Folketingsvalg – højresidet)

Vi skal selvfølgelig også betragte et højresidet binomialtest og denne gang ikke bare kigge på terningekast. Der er i den politiske verden en formodning om, at et politisk parti A er gået frem siden sidste valg, hvor partiet fik 22% af stemmerne. Derfor har man gennemført en meningsmåling med henblik på at give en vurdering af, om denne påstand kan have hold i virkeligheden. Man har udspurgt i alt 600 personer, hvoraf 158 fortalte, at de vil stemme på parti A næste gang, der er valg. Giv på baggrund af meningsmålingen en vurdering af, om forudsigelsen mon er rigtig, dvs. at partiet A har fået større tilslutning i befolkningen.

I de tidligere eksempler med terningekast var det oplagt, at betingelserne for at benytte et binomialtest var opfyldt. Fremover skal vi dog være omhyggelige med at godtgøre, at det er rimeligt at bruge et binomialtestet. For det første skal vi huske, at hypotesetest handler om, at man ud fra en stikprøve vil vurdere en egenskab for hele populationen. I dette tilfælde har vi klart:

Population:	Hele befolkningen
Stikprøve:	Den lille gruppe, der udspørges.
Binomialtest:	Ja, basiseksperimentet består i at udspørge 1 person. Der er to mulige udfald: Basiseksperimentet lykkes, hvis personen stemmer på parti A og mislykkes, hvis personen ikke stemmer på A. Der er næsten fast basissandsynlighed: Godt nok er eksperimentet "uden tilbagelægning", men da stikprøven er meget lille i forhold til befolkningen (populationen), vil sandsynlighederne ved hver udspørgning stort set være uændret. Udfaldene kan antages uafhængige, hvis man fx undgår at spørge personer fra samme familie og lignende. Bortset fra det, skal man selvfølgelig spørge bredt og tilfældigt i befolkningen.

Vores nulhypotese og alternative hypotese skal se således ud:

H_0 :	Tilslutningen til parti A er uændret siden sidste valg.
H_1 :	Tilslutningen til partiet A er vokset.

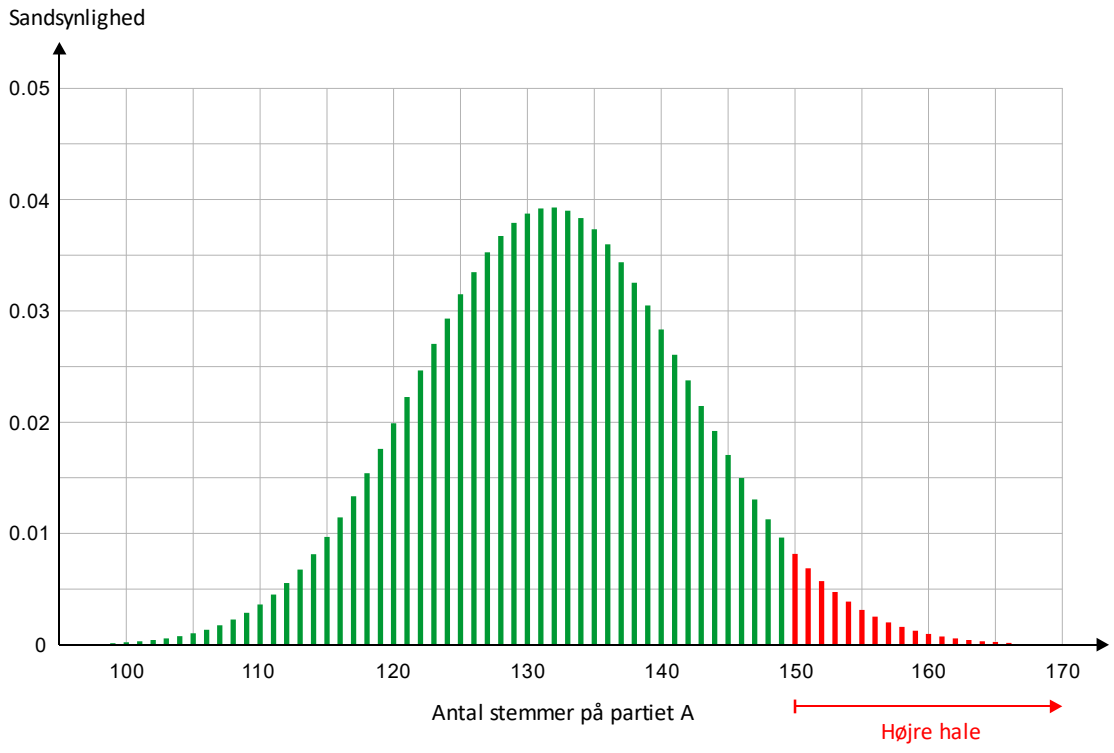
Vi foretager altså et *højresidet* test, eftersom vi kun har fokus på, om tilslutningen til parti A er *vokset*, ikke at den eventuelt måtte være faldet! Derfor har vi kun en kritisk hale i højre side. Igen benytter vi et signifikansniveau på 5%. Som sædvanlig antager vi, at nulhypotesen er rigtig. Dermed er basissandsynligheden (sandsynlighedsparameteren) lig med $p = 0{,}22$, antalsparameteren er $n = 600$ og teststørrelsen er 158. Vi søger det kritiske område. Denne hale skal have en samlet sandsynlighed på højst 5%. Vi prøver os lidt frem med forskellige kumulerede sandsynligheder og finder følgende:

$$\begin{aligned} bincdf(600, 0.22, 148) &= 0{,}9465254378 \\ bincdf(600, 0.22, 149) &= 0{,}9561651165 \end{aligned}$$

Vi kan ikke bruge 148 som højre endepunkt i acceptområdet, for så vil halen få en sandsynlighed på over 5%. 149 er den største værdi, som giver anledning til en hale, som højst har en sandsynlighed på 5%. Vi slutter, at højre endepunkt i acceptområdet er $k_h = 149$. Processen kan beskrives således: Man søger den *største* værdi af r, for hvilket den kumulerede frekvens $bincdf(600, 0.22, r)$ er *skarpt mindre* end 0,95. Herefter lægger man 1 til for at få højre endepunkt i acceptområdet, altså $k_h = 148 + 1 = 149$. Alt i alt fås et acceptområde på $\{0, \ldots, 149\}$. Ovenfor har der været lidt prøven sig frem. Man kan dog også benytte fraktiler for fordelingsfunktionen. Det skal vi dog ikke gå nærmere ind på. Nemmest er det naturligvis, hvis CAS-værktøjet direkte kan angive acceptområdet.

Eftersom teststørrelsen 158 *ikke* er med i acceptområdet, er den med i det kritiske område, hvorfor vi vælger at afvise nulhypotesen og godtage den alternative hypotese. Parti A har altså formentlig en øget tilslutning i befolkningen i forhold til sidste valg.

Nedenfor et pindediagram over sandsynlighedsfordelingen med indtegnet rød hale. Ikke alle værdier fra 0 til 600 er vist, da de har så små sandsynligheder, at det ikke kan ses.



Bemærkning 4.6

I et højresidet binomialtest fås det højre endepunkt k_h i acceptområdet på følgende måde: Lad $r_{\max}$ være den *største* værdi af r, for hvilket den kumulerede sandsynlighed er *skarpt mindre* end $1-\alpha$, altså $bincdf(n,p,r)<1-\alpha$. Da er $k_h = r_{\max}+1$. Acceptområdet er dermed $\{0,\ldots,k_h\}$.

□

Bemærkning 4.7

Det er vigtigt at forstå, at man *ikke* må foretage valget mellem et tosidet test, et venstresidet test og et højresidet test ud fra viden om stikprøven. Valget skal være foretaget *før* stikprøven tages, eller hvis den allerede foreligger, skal man se bort fra den i sine argumenter. Det samme gælder valget af signifikansniveau α.

□

Vi vil slutte dette afsnit af med en skabelon, som kan bruges som en rettesnor til at huske at overveje alle de vigtige punkter i forbindelse med en binomialtest. Her skal det for eksempel overvejes, om der overhovedet er tale om et binomialforsøg.

Binomialforsøg	
Er det binomialforsøg?	
Basiseksperiment	
Antalsparameter (n)	
Sandsynlighedsparameter (p)	
Succes	
Stokastisk variabel X	
Binomialtest	
Population	
Stikprøve	
Tosidet/enkeltsidet test?	
Signifikansniveau	
Teststørrelsen	

På næste side kan se, hvordan man for eksempel kunne udfylde den i forbindelse med eksempel 4.5 med tilslutningen til et politisk parti.

Binomialforsøg	
Er det binomialforsøg?	Ja, et basisforsøg udføres mange gange. Da undersøgelsen naturligvis er "uden tilbagelægning", så ændrer sandsynligheden sig i princippet en smule hver gang der spørges, men da stikprøven er meget mindre end populationen, kan vi se bort fra dette. Svarene fra hver person antages uafhængige.
Basiseksperiment	Én person udspørges om denne vil stemme på A eller ej
Antalsparameter (n)	600
Sandsynlighedsparameter (p)	0,22
Succes	Hvis personen vil stemme på partiet A
Stokastisk variabel X	Angiver antallet af personer, som vil stemme på A.
Binomialtest	
Population	Hele Danmarks befolkning
Stikprøve	De 600 personer, som udspørges.
Tosidet/enkeltsidet test?	Højresidet
Signifikansniveau	0,05
Teststørrelsen	158 vil stemme på partiet A

4.3 Binomialtest: Vigtige grundlæggende erkendelser

Det er vigtigt at bemærke, at resultatet af et binomialtest – eller et hypotesetest i al almindelighed – afhænger af stikprøven. Man kan således sagtens forestille sig, at man foretager en stikprøve og accepterer nulhypotesen på grundlag af den, mens man ved en anden stikprøve havde forkastet nulhypotesen. Blandt andet derfor skal man passe på med at blive for kategorisk med sine konklusioner ud fra Binomialtesten. Hvis binomialtesten viser, at teststørrelsen er i acceptområdet, så må man for eksempel *ikke* sige, at så er nulhypotesen *rigtig*. Man kan sige: "vi *accepterer* nulhypotesen" eller "vi kan *ikke forkaste* nulhypotesen". Tilsvarende hvis teststørrelsen er i det kritiske område: Man må *ikke* sige, at nulhypotesen er *forkert*. Man kan sige: "Vi *afviser* nulhypotesen" eller "vi *forkaster* nulhypotesen".

Desuden afhænger konklusionen af valget af signifikansniveau α. Der er altså noget subjektivt i spil her. Det er dog meget almindeligt at anvende 5% som signifikansniveau. Der kan også undertiden ligge noget subjektivt i valget af, om testet skal være tosidet eller ensidet. Vi så i eksempel 4.1 og 4.3, at det kan få betydning for konklusionen.

Husk at en stikprøve skal foretages, så den er *tilfældig* og *repræsentativ*. Man skal være opmærksom på, at stikprøven ikke får en *skævhed* (har bias). I tilfældet med en folketingsprognose er det for eksempel vigtigt, at der spørges tilfældigt og bredt i Danmark.

Størrelsen af stikprøven, altså værdien af n, har klart en betydning. Hvis stikprøven er tilfældig og repræsentativ, vil resultatet normalt blive mere troværdigt, jo flere der spørges. Vi ser på problematikken i et eksempel senere i dette afsnit. Alt dette indikerer, at man ikke altid kan forvente at få det rigtige svar. Det leder naturlig frem til følgende:

Type I og type II fejl

- også undertiden kaldet fejl af 1. art og 2. art. Fejl af type I, hvor en sand hypotese forkastes og fejl af type II, hvor en forkert hypotese accepteres. Man kan sammenligne det lidt med situationen på et hospital, hvor der tages en prøve fra en person med henblik på at klarlægge, om personen er smittet med en bestemt bakterie eller ej. Desværre kan man ikke foretage fuldstændig sikre afgørelser i medicinske undersøgelser. Man opererer med begreberne *falske negative* og *falske positive* testresultater. En *falsk negativ* hentyder til, at prøven viser, at personen ikke er smittet, men at vedkommende i virkeligheden er smittet – altså en type I fejl. En *falsk positiv* hentyder til, at prøven viser, at personen er smittet, men at denne i virkeligheden ikke er det – en type II fejl. I dette tilfælde vil man normalt sige, at den af type I er den værste fejl. Skematisk kan opsummeres således:

	Nulhypotesen er sand	Nulhypotesen er forkert
Nulhypotesen accepteres	Korrekt afgørelse	Type II fejl
Nulhypotesen afvises	Type I fejl	Korrekt afgørelse

Men hvad er sandsynlighederne for type I og type II fejl? Ja i tilfældet med type I fejl er det nemt at besvare: Den er lig med signifikansniveauet α, eller rettere højst α, på grund af den diskrete fordeling. Hvis nulhypotesen er sand, så har vi jo netop valgt at afvise nulhypotesen, hvis teststørrelsen ligger i det kritiske område af størrelse (højst) α. Man kunne så forestille sig folk forslå, at så kan man bare sætte signifikansniveauet ned! Men dette er ikke nødvendigvis fornuftigt, for så er der en større sandsynlighed for at begå en fejl af Type II – det er en *trade-off* situation. Hvis man sætter signifikansniveauet ned, så gør man det sværere at afvise en sand nulhypotese. Til gengæld bliver det lettere at komme til at acceptere en nulhypotese, som burde afvises, altså begå en type II fejl!

Man kan ikke uden videre udregne sandsynligheden for en fejl af type II, for i den alternative hypotese er der ikke antaget én bestemt sandsynlighed. I nulhypotesen har man altid en hypotese af typen $p = p_0$, mens den alternative hypotese er en af typerne $p \neq p_0$, $p > p_0$ eller $p < p_0$. Sandsynlighed for en fejl at type II afhænger således af, hvilken værdi for p, man måtte vælge som udgangspunkt. Vi ser på det i opgave 409. Ellers vil vi ikke gå yderligere ind i denne problematik.

For en ordens skyld skal det nævnes, at nogle forfattere i forbindelse med et højresidet tests vælger at skrive den nulhypotesen som $p \le p_0$, fremfor $p = p_0$, men der er ikke reelt nogen forskel i beregningerne, for alle beregninger går ud fra, at $p = p_0$. Tilsvarende kalder disse forfattere nulhypotesen for $p \ge p_0$ i tilfældet med et venstresidet tests. I et højresidet test er det kun værdier i højre side, som kan forkaste nulhypotesen, mens det kun er værdier i venstre side, som kan forkaste nulhypotesen for et venstresidet test.

Det er klart, at hvis man opnår en meget bemærkelsesværdig værdi for teststørrelsen, så bør det overvejes, om man skal tage en ny stikprøve som en ekstra sikkerhed. En anden mulighed er at opstille en helt ny hypotese og derefter teste den mod en helt ny stikprøve.

Lad os endelig betragte et eksempel, som tydeligt illustrerer den betydning, som stikprøvestørrelsen *n* har i et binomialtest.

Eksempel 4.8 (Stikprøvestørrelsens betydning)

Man er i tvivl om, hvorvidt en mønt er ægte, dvs. om frekvensen af plat i det lange løb er er 50%. Spørgsmålet ønskes afgjort eller vurderet på baggrund af en stikprøve og en tosidet binomialtest med et signifikansniveau på 5%.

Stikprøve 1: Mønten kastes 10 gange og 2 gange viser mønten plat.
Stikprøve 2: Mønten kastes 100 gange og 20 gange viser mønten plat.

Det er klart, at vi har at gøre med binomialforsøg, og at basissandsynligheden er 1/2. Nulhypotesen er, at mønten er ægte, dvs. at $p = \frac{1}{2}$, mens den alternative hypotese er, at mønten er uægte og at $p \ne \frac{1}{2}$. Uden at gå i detaljer med udregningerne giver vi her uden videre resultaterne:

Binomialtesten på baggrund af stikprøve 1

Nulhypotesen *accepteres*, fordi teststørrelsen på 2 plat ligger i acceptområdet.

Binomialtesten på baggrund af stikprøve 2

Nulhypotesen *forkastes*, fordi teststørrelsen på 20 plat ligger i det kritiske område.

Man kan stille sig det spørgsmål, hvorfor resultatet af de to binomialtest ikke giver det samme nu, da tallene i stikprøve 2 bare er ganget op med 10? Er det virkelig mere usandsynligt, at få 20 plat i 100 kast end 2 plat i 10 kast, hvis nulhypotesen er sand? Ja det er det! Det er *de store tals lov*, som er i spil. Oversat til denne situation siger loven, at jo flere gange vi slår med mønten, jo mere sandsynligt er det, at frekvensen af plat for den antaget ægte mønt er tæt på $\frac{1}{2}$. Faktisk skal man helt op på 40 plat i stikprøve 2, for at man vil acceptere nulhypotesen! Eksemplet indikerer også, at jo større stikprøver, man tager (underforstået under tilfældige og repræsentative forhold) jo mindre procentvis spredning vil der være fra den antaget sande middelværdi på $\frac{1}{2}$. Samtidigt indser vi den meget vigtige pointe, at binomialtesten *ikke* kan bruges til at konkludere noget udelukkende på baggrund af *procenter*! Man skal kende de konkrete *antal*!

4.4 Binomialtest via *p*-værdien

I dette afsnit skal vi kort omtale en variant til den type binomialtests, som blev gennemgået tidligere i dette kapitel. I stedet for at bestemme et acceptområde og et kritisk område med udgangspunkt i nulhypotesen, kan man i stedet bestemme sandsynligheden for, at den binomialfordelte stokastiske variabel X antager en værdi, som er lige så ekstrem eller mere ekstrem end den observerede værdi (teststørrelsen) – igen antaget at nulhypotesen er sand. Hvis denne sandsynlighed, kaldet p-værdien, er $\leq \alpha$ (signifikansniveauet), så forkastes nulhypotesen, ellers accepteres den. Bemærk, at denne p-værdi *ikke* har noget med sandsynlighedsparameteren p i binomialfordelingen at gøre!

Definition 4.9 (p-værdi)

Til teststørrelsen (den observerede værdi af X) knyttes den såkaldte p-værdi, som angiver sandsynligheden for at den binomialfordelte stokastiske variabel X antager en værdi, som er *lige så ekstrem eller mere ekstrem* end den pågældende teststørrelse – under antagelse at nulhypotesen er sand.

Eksempel 4.10

Lad os se, hvordan det ville se ud i situationen i eksempel 4.1, hvor vi opstillede en tosidet test for på et 5% signifikansniveau at afgøre, om en terning var ægte eller ej. Der blev kastet 48 gange med en terning og det gav i alt 3 seksere. Nulhypotesen er, at terningen er ægte. Med den antagelse har vi altså en binomialfordelt stokastisk variabel X med antalsparameter $n = 48$ og sandsynlighedsparameter $p = \frac{1}{6}$. Den observerede værdi for X er 3 (teststørrelsen). Vi vil forvente $\mu = n \cdot p = 48 \cdot \frac{1}{6} = 8$ seksere i middel. Teststørrelsen er altså i den lave ende. Mere ekstreme værdier for X ville være 0, 1 eller 2. Dermed kan vi bestemme sandsynligheden for at X i venstre side antager en værdi, som er mindst lige så ekstrem som 3. Det er den kumulerede sandsynlighed, som vi her betegner med *bincdf*, ligesom i bemærkning 4.2.

$$bincdf(48, \tfrac{1}{6}, 3) \;=\; 0{,}030711$$

Men der er også mulighed for, at X kan antage ekstremt store værdier. I tosidede tests vælger man at tage højde for det ved at gange p-værdien med 2. I dette tilfælde bliver p-værdien altså $2 \cdot 0{,}030711 = 0{,}061422$. Da p-værdien er større end signifikansniveauet på 5%, accepterer vi nulhypotesen. Altså samme konklusion som i eksempel 4.1.

□

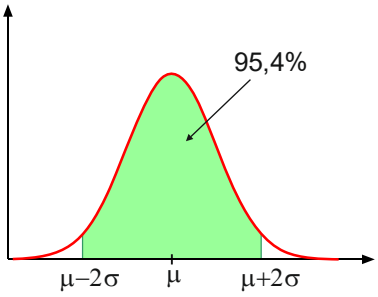
Bemærkning 4.11

I tilfældet med enkeltsidet test, skal man naturligvis ikke gange med 2. Her er p-værdien blot sandsynligheden for den hale, som starter med teststørrelsen og går enten mod venstre eller højre, alt efter om teststørrelsen er mindre end eller større end middelværdien μ.

4.5 Konfidensinterval for andel

I de forrige afsnit i kapitel 4 har vi set, hvorledes man kan benytte et binomialtest til at acceptere eller forkaste en nulhypotese af typen $p = p_0$. Her er altså tale om, at man lægger sig fast på en bestemt sandsynlighed p_0 som udgangspunkt for testet. Lad os sige, at vi accepterede nulhypotesen. Man kan hurtigt overbevise sig om, at hvis man havde ændret værdien for p_0 en smule og gentaget testet med det nye udgangspunkt, så kan det sagtens være, at man også ville komme frem til at acceptere den nye nulhypotese. Så endnu en gang ser vi, at binomialtest ikke fører til en entydig sand konklusion. Det ville være rart, hvis man kunne *kvantificere* usikkerheden på p_0. Det er her begrebet konfidensinterval kommer ind.

Vi ved, at normalfordelingen under visse betingelser er en god tilnærmelse til binomialfordelingen (afsnit 3.3). Vi skal desuden udnytte følgende egenskab for en normalfordelt stokastisk variabel X med middelværdi μ og spredning σ (jf. opgave 327):



(1) $$P(\mu - 2\sigma \le X \le \mu + 2\sigma) \;=\; 95{,}4\%$$

For en binomialfordelt stokastisk variabel X gælder ifølge sætning 3.6, at middelværdi og spredning er givet ved $\mu = E(X) = n \cdot p$ og $\sigma(X) = \sqrt{n \cdot p \cdot (1-p)}$. Dermed bliver (1) til:

$$P\left(n \cdot p - 2 \cdot \sqrt{n \cdot p \cdot (1-p)} \le X \le n \cdot p + 2 \cdot \sqrt{n \cdot p \cdot (1-p)}\right) \approx 0{,}95$$

Dividerer vi med n i dobbeltuligheden fås:

$$P\left(p - 2 \cdot \sqrt{\frac{p \cdot (1-p)}{n}} \le \frac{X}{n} \le p + 2 \cdot \sqrt{\frac{p \cdot (1-p)}{n}}\right) \approx 0{,}95$$

I udtrykket for spredningen $\sqrt{p \cdot (1-p)/n}$ indgår p, som vi ikke kender. Vi tillader os at udskifte den med estimatoren $\hat{p} = X/n$, hvorved vi får:

$$P\left(p - 2 \cdot \sqrt{\frac{X/n \cdot (1-X/n)}{n}} \le \frac{X}{n} \le p + 2 \cdot \sqrt{\frac{X/n \cdot (1-X/n)}{n}}\right) \approx 0{,}95$$

Uligheden i parentesen kan omskrives, så den ukendte andel p står i midten:

$$P\left(X/n - 2 \cdot \sqrt{\frac{X/n \cdot (1-X/n)}{n}} \le p \le X/n + 2 \cdot \sqrt{\frac{X/n \cdot (1-X/n)}{n}}\right) \approx 0{,}95$$

For en konkret stikprøve, hvor basiseksperimentet lykkes k gange, dvs. $X = k$, fås estimatet $\hat{p} = \frac{k}{n}$ for den ukendte andel p, og ovenstående viser, at intervallet

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

i ca. 95% af tilfældene vil indeholde den rigtige værdi p for andelen. Ovenstående er gennemført meget løst og hurtigt, og $\approx$ gælder kun, når n og p opfylder visse betingelser.

Sætning 4.12 (Konfidensinterval for en andel)

Givet en stokastisk variabel X, som er binomialfordelt med kendt antalsparameter n, men ukendt sandsynlighedsparameter p. Antag udtaget en stikprøve af størrelse n, hvor basiseksperimentet lykkes k gange. Da er $\hat{p} = k/n$ et estimat på den rigtige sandsynlighed p. Usikkerheden på $\hat{p}$ er givet ved et 95%-konfidensinterval:

$$(4) \qquad \left[\hat{p} - 2 \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}},\ \hat{p} + 2 \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}} \right]$$

Formel (4) gælder ikke for små stikprøver. Som en tommelfingerregel skal følgende være opfyldt, hvis tilnærmelsen skal være god: $n \geq 30 \ \wedge \ n \geq \frac{10}{\hat{p}} \ \wedge \ n > \frac{10}{1-\hat{p}}$.

Med betegnelsen "95%-konfidensinterval" menes følgende: Det er klart, at intervallet afhænger af stikprøven, mens den rigtige værdi p er fast. Hvis man tog et meget stort antal stikprøver på tilfældig vis, så ville ca. 95% af disse intervaller indeholde den korrekte værdi p. Vi ser på problematikken i eksempel 4.15 på næste side.

Eksempel 4.13 (Meningsmåling genbesøgt)

Vi vil kigge på situationen i eksempel 4.5, blot med den forskel, at vi har en formodning om, at partiet har *uændret* tilslutning i forhold til sidste valg. hvor resultatet blev 22%. I stedet for at lave et tosidet test, vil vi kigge på situationen fra den nye synsvinkel med konfidensinterval. Stikprøvestørrelsen var på 600 og der blev registreret 158, som ville stemme på partiet. Dermed kan vi bestemme et estimat for den ukendte andel p.

$$\hat{p} = \frac{k}{n} = \frac{158}{600} = 0{,}2633$$

Lad os kontrollere betingelsen for at bruge (4): $10/\hat{p} = 38{,}0$ og $10/(1-\hat{p}) = 13{,}6$. Med $n = 600$ er de rigeligt opfyldt. Den statistiske usikkerhed er:

$$2 \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}} = 2 \cdot \sqrt{\frac{0{,}2633 \cdot (1-0{,}2633)}{600}} = 0{,}0360$$

Dermed haves 95%-konfidensintervallet:

$$[0{,}2633 - 0{,}0360;\ 0{,}2633 + 0{,}0360] \text{ eller } [0{,}227;\ 0{,}299]$$

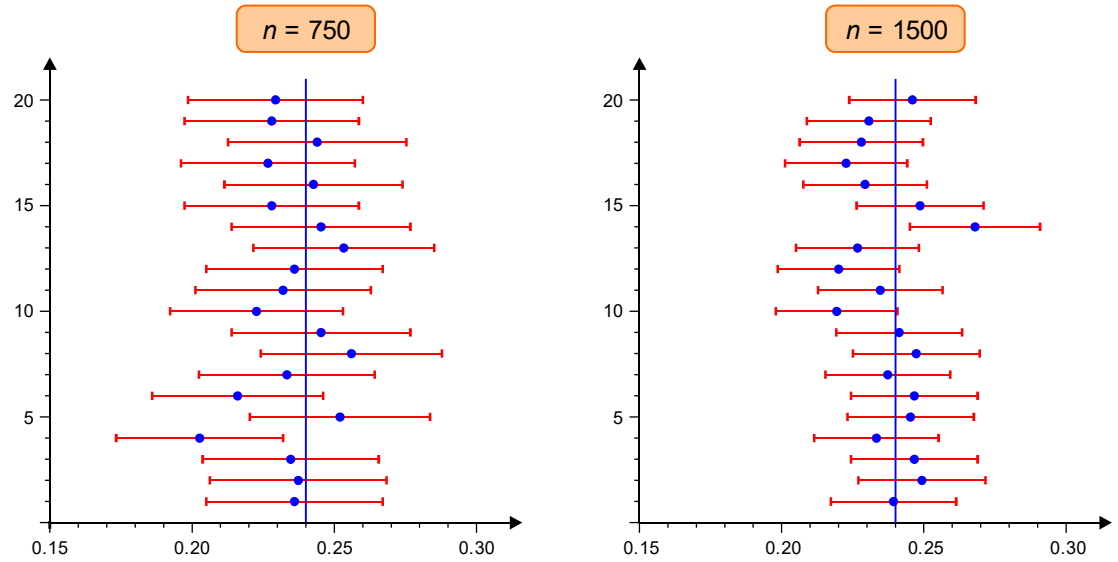
Et interval, som strækker sig fra 22,7% til 29,9%. Dette interval er nu fast og vil enten indeholde den ukendte og sande sandsynlighed eller ej. Tager man et meget stort antal stikprøver, vil ca. 95% af alle stikprøverne have et tilhørende konfidensinterval, som indeholder den sande værdi for p. Det foreliggende konfidensinterval indeholder *ikke* sandsynligheden på 22% fra nulhypotesen. Derfor vælger vi at forkaste nulhypotesen om at 22% er den nuværende sande andel, som stemmer på partiet A.

Bemærkning 4.14 (Binomialtest vs konfidensinterval for andel)

Der er en approksimativ ækvivalens mellem et binomialtest og et konfidensinterval for en andel: *En teststørrelse k befinder sig i acceptområdet i et binomialtest for nulhypotesen* $p = p_0$ *hvis og kun hvis* p_0 *befinder sig i konfidensintervallet hørende til* $\hat{p} = k/n$. Det gælder kun approksimativt, da der kan være små "randeffekter" som skyldes, at binomialfordelingen jo er approksimeret med en normalfordeling. Man kan vise, at højre- og venstresidede binomialtest (5%) også har deres tilhørende 95% konfidensintervaller, nemlig henholdsvis: $[\hat{p} - \Delta p, 1]$ og $[0, \hat{p} + \Delta p]$, hvor $\Delta p = 1{,}65 \cdot \sqrt{\hat{p} \cdot (1 - \hat{p})/n}$.

Eksempel 4.15 (Simulation af konfidensintervaller)

Man kan simulere konfidensintervaller på en computer: Vi antager, at vi kender den (korrekte) stemme-andel for et parti i hele befolkningen, her $p = 0{,}24$. Så udtrækker vi gentagne gange og på tilfældig vis en stikprøve på 750 for at se hvilken estimeret stemmeandel, det i hvert tilfælde giver for partiet. Venstre del af figuren nedenfor viser resultatet af 20 simulerede stikprøver. For hver enkelt stikprøve er den estimerede stemmeandel angivet ved en blå prik samt et konfidensinterval beregnet efter formlen i sætning 4.12. Den lodrette blå linje angiver den korrekte stemmeandel på 24%. For det første ser vi, at der er en del variation. Nogle gange er estimatet en del under den korrekte værdi, andre gange væsentligt over eller tæt på. Hvad der er interessant er, at de fleste 95%-konfidensintervaller rammer den blå linje. Her faktisk præcist 95%. Et enkelt interval rammer ikke. Det behøver nu ikke altid være præcist 95%, der rammer. Nogle gange vil det måske være 85% eller 100%. Simulerer man en stikprøve rigtig mange gange, fx 100000 gange, så kan man dog være temmelig sikker på, at man får et resultat tæt på de 95%.



Den højre del af figuren ovenfor illustrer resultatet efter at have simuleret en stikprøve 20 gange, hvor stikprøvestørrelsen er dobbelt så stor, dvs. 1500. Som det forventes af (4) bliver konfidensintervallerne smallere. Spredningen i stikprøveresultater er blevet mindre. Samtidigt vil de estimerede stemmeandele også gennemgående ligge tætter på den korrekte. Så jo større stikprøver, jo større nøjagtighed – normalt.

Bemærkning 4.16 (Simuleringsprogram)

Med de fleste CAS-værktøj kan man simulere, hvor stor en del af konfidensintervallerne som rammer den rigtige værdi for p. Det gælder om at skrive en programstump, hvor man på forhånd sætter n og p til konkrete værdier samt vælger et antal stikprøver N. Derefter laver man en løkke, som gennemløbes N gange. I hvert gennemløb genereres en stikprøve bestående af ét element fra en binomialfordelt stokastisk variabel. For hver af disse udregnes konfidensintervallets grænser og sammenlignes med den rigtige værdi af p …

4.6 Konfidensinterval for en middelværdi

Indholdet af dette afsnit er ikke obligatorisk i gymnasiet i dag. Årsagen til, at det alligevel er medtaget er, *konfidensinterval for en middelværdi* eller et gennemsnit er et vigtigt emne i statistik. Det er mere generelt end tilfældet med konfidensinterval for en andel, som vi lige har kigget på. Den sætning, som vi kommer til at fremsætte, kan faktisk bruges til at bestemme et konfidensinterval for en andel, om end på en lidt anden måde. Udgangspunktet for den statistiske analyse er, at man har en *simpel tilfældig stikprøve*, hvor den *i*'te observation abstrakt set kan beskrives ved en stokastisk variabel X_i. "Simpel og tilfældig" går på, at observationerne er uafhængige, og at de stokastiske variable har samme fordeling. Resultatet af en konkret stikprøve kan være $\{x_1, x_2, \ldots, x_n\}$, hvor de enkelte værdier x_i er de værdier, som de stokastiske variable antager i det konkrete tilfælde: $X_i = x_i$. Det kunne være, at stikprøven bestod i at fange en række fisk og måle deres længde, før man lod dem slippe løs igen. En konkret stikprøve kunne da være på formen $\{23, 41, 34, 38, 42, 27, \ldots, 36\}$ regnet i cm. Tog vi en ny stikprøve, ville det sædvanligvis give en helt anden række af observationer. Her er det så vigtigt, at vi tager fat i de stokastiske variable X_i, som jo indeholder informationen om, hvordan de enkelte observationer kan variere. Vi antager som sagt, at de stokastiske variable har samme fordeling og er uafhængige. En oplagt størrelse at se på er *stikprøvegennemsnittet*, som er defineret ved:

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} \tag{5}$$

Det er en ny stokastisk variabel, også kaldet en *estimator* for middelværdien. Har man værdier fra en konkret stikprøve, vil den stokastiske variable $\overline{X}$ som værdi $\overline{x}$ have gennemsnittet af værdierne i den konkrete stikprøve. Altså en *middelværdi*! I tilfældet med stikprøven af fisk ville det være:

$$\overline{x} = \frac{23 + 41 + 34 + 38 + 42 + 27 + \ldots + 36}{88} = 36{,}15$$

hvis der altså var 88 fisk i stikprøven. Vi konstaterer, at fiskenes gennemsnitslængde i stikprøven er 36,2 cm. Men det var kun én stikprøve. Havde vi foretaget en ny stikprøve, ville det sandsynligvis have givet noget andet. Og vi skal huske, at vi jo gerne vil udtale os om populationen ud fra stikprøven! Det er jo ikke sikkert, at gennemsnitslængden for fiskene i hele søen er 36,2 cm. Så hvor meget kan vi stole på stikprøve-resultaterne? Det er her, at variationen beskrevet ved de stokastiske variable X_i kommer ind i billedet! Det er så også her, at det bliver svært. Man har følgende sætning:

Sætning 4.17 (Den centrale grænseværdisætning)

Givet en simpel tilfældig stikprøve bestående af n elementer, beskrevet ved de stokastiske variable $X_1, X_2, \ldots, X_n$, hvor $E(X_i) = \mu$ og $Var(X_i) = \sigma^2$ for alle i. Da vil stikprøvegennemsnittet $\overline{X}$ være approksimativt normalfordelt med middelværdi μ og varians σ^2/n, når blot stikprøven er stor.

Den centrale grænseværdisætning er en hjørnesten i sandsynlighedsregningen. Et dybt og overraskende resultat. Overraskende, da stikprøvegennemsnittet ender med at blive ca. normalfordelt uagtet, at fordelingen af X_i kan være en hel anden! For at approksimationen er god, kræver det dog en tilstrækkelig stor stikprøve. En anden god ting er, at middelværdien af stikprøvegennemsnittet (betragtet som stokastisk variabel) er den samme som middelværdien af X_i. Og en anden god nyhed: Variansen af stikprøvegennemsnittet er lig med variansen af hver X_i *divideret med n*! Spredningen er altså blevet mindre af, at vi har taget gennemsnittet. Passer meget godt med vores intuition fra fysikforsøg!

Sætning 4.18 (Konfidensinterval for middelværdien – det generelle tilfælde)

Givet en simpel tilfældig stikprøve bestående af n elementer: $\{x_1, x_2, \ldots, x_n\}$. Et estimat for den sande middelværdi μ i populationen er den *empiriske middelværdi* af værdierne i stikprøven:

(6) $$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Et estimat for den korrekte spredning σ (på en enkelt størrelse) i populationen er *stikprøvespredningen s* givet ved

(7) $$s = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n-1}}$$

Stikprøvevariansen er tilsvarende s^2. Den såkaldte *standardfejl* på gennemsnittet fås herefter ved at dividere med $\sqrt{n}$:

(8) $$s_{\overline{x}} = \frac{s}{\sqrt{n}} \quad \Leftrightarrow \quad s_{\overline{x}}^2 = \frac{s^2}{n}$$

Et *approksimativt* 95%-konfidensinterval for middelværdien er da følgende:

(9) $$\left[\overline{x} - 2 \cdot s_{\overline{x}},\ \overline{x} + 2 \cdot s_{\overline{x}}\right] \quad \text{eller} \quad \left[\overline{x} - 2 \cdot \frac{s}{\sqrt{n}},\ \overline{x} + 2 \cdot \frac{s}{\sqrt{n}}\right]$$

Det betyder, at sandsynligheden for, at et sådant interval – som jo afhænger af stikprøven – indeholder den korrekte værdi for middelværdien μ, er ca. 95%.

En tommelfingerregel er, at n skal opfylde $n \geq 30$ for at approksimationen er tilstrækkelig god.

Vi skal ikke bevise sætningen, da den igen beror på den komplicerede centrale grænseværdisætning. Igen bemærker vi som forventet, at jo større stikprøven er, jo mindre bliver standardfejlen på middelværdien.

Bemærkning 4.19 (Konfidensinterval for en andel)

Vi skal se, at sætning 4.18 kan bruges til at bestemme et konfidensinterval for en andel. For det første indfører vi hurtigt den såkaldte *Bernouilli fordeling*, som er en diskret fordeling, som kun kan antage to værdier, nemlig 1 med sandsynlighed p og 0 med sandsynlighed $1-p$. Reelt svarer det blot til en binomialfordeling med antalsparameter $n=1$. På den måde er et Bernouilli-forsøg det samme som et basiseksperiment i binomialfordelingen. Lad os sige, at vi har at gøre med en meningsmåling, hvor man spørger hver person, om denne stemmer på partiet A eller ej. Da kan man lade den Bernoulli-fordelte stokastiske variabel X_i være lig med 1, hvis den i'te udspurgte person vil stemme på partiet A, og 0, hvis denne ikke vil stemme på partiet. På den måde tæller summen af X_i'erne (altså $X_1+X_2+\ldots+X_n$) op, hvor mange der i stikprøven stemmer på partiet A. Når man så efterfølgende dividerer med n, får man stikprøvegennemsnittet (5). Det vil være et estimat for den andel, som stemmer på partiet i hele populationen. Sætning 4.18 kan da bruges til at give et konfidensinterval for middelværdien. Det involverer udregning af stikprøvespredningen (7). Med andre ord en lidt anden procedure. Det skal dog nævnes, at konfidensintervallet i sætning 4.12 er at foretrække, når man skal bestemme konfidensintervaller for andele.

Eksempel 4.20 (Pakkebude)

Et firma, som distribuerer pakker til kunder, er interesseret i at danne sig et overblik over, hvor mange pakker en afdeling med pakkebude kan håndtere i gennemsnit hver dag. Firmaet foretager derfor en stikprøve 50 dage spredt ud over året. Det gav følgende værdier for det daglige antal pakker:

339, 317, 529, 559, 412, 398, 486, 369, 549, 479, 354, 311, 483, 405, 470, 555, 431, 575, 387, 471, 396, 406, 345, 425, 376, 448, 581, 357, 326, 509, 351, 302, 504, 415, 423, 329, 611, 446, 533, 505, 405, 216, 411, 549, 432, 331, 411, 451, 382, 473

Vi ønsker at bestemme middelværdien samt et 95% konfidensinterval for middelværdien. Ved brug af sætning 4.18 fås først middelværdien:

$$\overline{x} = \frac{339+317+\ldots+473}{50} = 430{,}56$$

og derefter stikprøvespredningen:

$$s = \sqrt{\frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\ldots+(x_n-\overline{x})^2}{n-1}}$$
$$= \sqrt{\frac{(339-430{,}56)^2+(317-430{,}56)^2+\ldots+(437-430{,}56)^2}{50-1}} = 85{,}516$$

og standardfejlen på gennemsnittet:

$$s_{\overline{x}} = \frac{s}{\sqrt{n}} = \frac{85{,}516}{\sqrt{50}} = 12{,}094$$

som giver følgende 95% konfidensinterval for middelværdien:

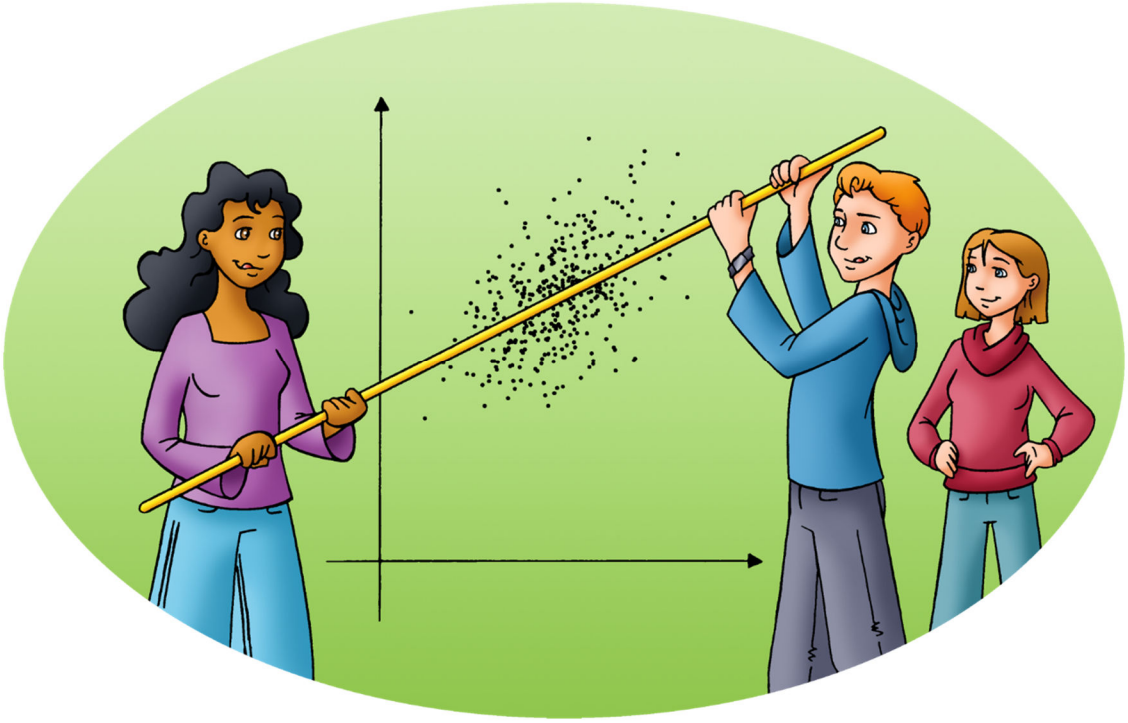
$$\left[430{,}56-2\cdot 12{,}094;\ 430{,}56+2\cdot 12{,}094\right] \quad \text{eller} \quad \left[406{,}4;\ 454{,}7\right]$$

Et godt estimat for, hvor meget pakkeafdelingen udbringer i gennemsnit pr. dag, er altså ca. 431 pakker. Konfidensintervallet indeholder med ca. 95% sandsynlighed den *korrekte* værdi for gennemsnittet, altså det gennemsnit, der vil være over lange perioder (populationen). Disse resultater afhænger selvfølgelig af, at stikprøven er repræsentativ, og at der ikke pludseligt indtræder nye regler eller der forekommer ændringer i mandskabet m.m.

□

Ellers henvises læseren til projektopgaven 419, hvor man skal estimere gennemsnitslængden for fisk i en sø – med tilhørende konfidensinterval. Her vil man også kunne studere, hvad man gør i tilfældet med små stikprøver.

5. Lineær regressionsanalyse



5.1 Indledning

Du har formodentligt allerede stiftet bekendtskab med *lineær regression* en hel del på nuværende tidspunkt, måske ved givet data at bruge en kommando eller trykke på en knap i et CAS-værktøj. I dette kapitel skal vi imidlertid studere begrebet mere indgående.

Regressionsanalyse er en statistisk metode, hvor man undersøger sammenhængen mellem en *forklarende variabel* x og en *responsvariabel* y. Regressionsanalyse har sit udspring i den sidste halvdel af 1800-tallet, hvor den britiske statistiker *Sir Francis Galton* (1822-1911) studerede sammenhængen mellem forældrenes højde og deres børns højde. I den særlige opgave 516 ser vi på problemstillingen. Her skal vi også præsenteres for en forklaring på betegnelsen *regression*.

I afsnit 5.2 skal vi se på lineær regression alene defineret ud fra en række datapunkter uden speciel struktur. I afsnit 5.3 skal vi derefter se på en udvidet lineær regressionsmodel, som involverer normalfordelingen.

5.2 Lineær regression og mindste kvadraters metode

Lad os sige, at der er givet n datapunkter $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, og at vi ønsker at bestemme den lineære funktion $f(x) = a \cdot x + b$, som er det *bedste fit* til datapunkterne. Da må vi først gøre os klart, hvad vi mener med "det bedste fit". Problemet blev studeret i begyndelsen af 1800-tallet, og det udmøntede sig i den metode, som i dag går under betegnelsen *mindste kvadraters metode* (engelsk: *Least Square Method*): Man søger at bestemme de værdier af a og b, som minimerer *kvadratsummen* af forskellen mellem datapunkternes y-værdier og funktionsværdierne i de tilhørende x-værdier:

$$(1) \qquad L = \sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2 = (y_1 - f(x_1))^2 + (y_2 - f(x_2))^2 + \ \ldots \ + (y_n - f(x_n))^2$$

Vi skal snart se, hvad det betyder geometrisk set.

Sætning 5.1 (Lineær regression – mindste kvadraters metode)

Givet n datapunkter $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Den linje $y = a \cdot x + b$, som minimerer størrelsen L i (1) har hældningskoefficient givet ved

$$(2) \qquad a = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

og konstantled givet ved

$$(3) \qquad b = \overline{y} - a \cdot \overline{x}$$

Bevis: Vi skal ikke bevise denne sætning. □

Bemærkning 5.2

Det er en interessant kendsgerning, at punktet $(\overline{x},\overline{y})$ ligger på regressionslinjen! Det følger direkte af (3): $b=\overline{y}-a\cdot\overline{x} \Leftrightarrow \overline{y}=a\cdot\overline{x}+b$.

□

Bemærkning 5.3

Udtrykket for hældningskoefficienten for regressionslinjen i sætning 5.1 kan alternativt skrives således:

$$a = \frac{\overline{x\cdot y}-\overline{x}\cdot\overline{y}}{\overline{x^2}-\overline{x}^2} \tag{4}$$

hvori der indgår flere middelværdier:

$$\overline{x\cdot y} = \frac{1}{n}\sum_{i=1}^{n} x_i\cdot y_i\,,\quad \overline{x} = \frac{1}{n}\cdot\sum_{i=1}^{n} x_i\,,\quad \overline{y} = \frac{1}{n}\cdot\sum_{i=1}^{n} y_i\,,\quad \overline{x^2} = \frac{1}{n}\cdot\sum_{i=1}^{n} x_i^2 \tag{5}$$

For at se et bevis for dette henvises interesserede læsere til mit tillæg *Lineær regression, kvadratsummer og forklaringsgrad* (se [L2]).

□

Definition 5.4 (Forklaringsgraden)

Givet følgende sæt af datapunkter: $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$. *Forklaringsgraden* R^2 for den regressionslinje, som hører til datapunkterne, er defineret ved:

$$R^2 = 1 - \frac{\sum_{n=1}^{n}\left(y_i-f(x_i)\right)^2}{\sum_{n=1}^{n}\left(y_i-\overline{y}\right)^2} \tag{6}$$

hvor f er den lineære funktion, som fremkommer ved lineær regression.

Bemærkning 5.5

Der findes flere formler for forklaringsgraden. To af dem er følgende:

$$R^2 = a^2\cdot\frac{\sum_{n=1}^{n}(x_i-\overline{x})^2}{\sum_{n=1}^{n}(y_i-\overline{y})^2} = a^2\cdot\frac{\overline{x^2}-\overline{x}^2}{\overline{y^2}-\overline{y}^2} \tag{7}$$

hvor a er hældningskoefficienten (2) for regressionslinjen. Vi skal ikke bevise dem. Den interesserede læser henvises igen til mit tillæg i [L2].

□

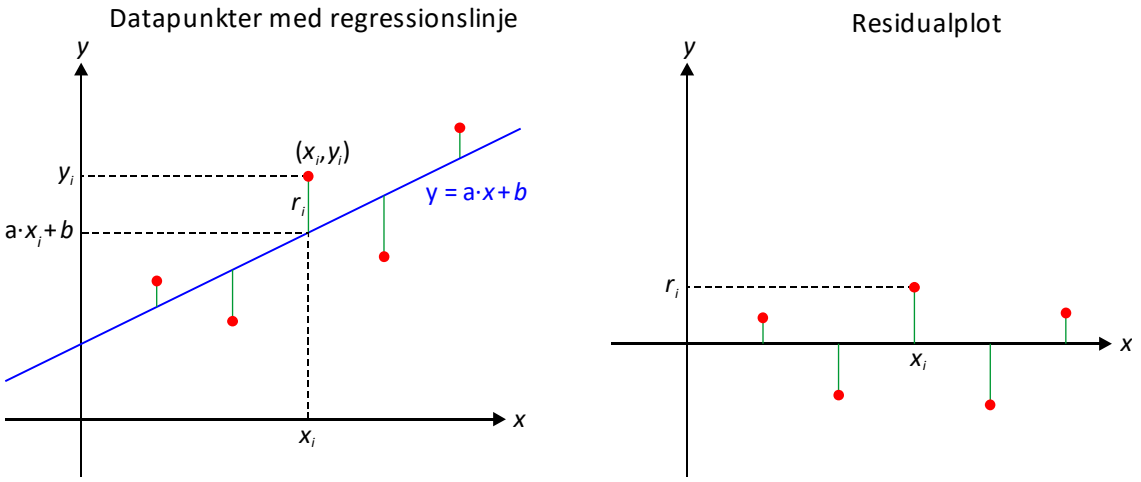
Man kan vise, at forklaringsgraden altid er et tal mellem 0 og 1: $0\le R^2\le 1$. Der gælder desuden, at $R^2=1$ hvis og kun hvis alle datapunkterne ligger præcist på linje. Overvej,

hvorfor det umiddelbart fremgår af definitionen (6)! Man ser også af (6), at R^2 er tæt på 1 netop når datapunkterne ligger tæt på regressionslinjen. Den kendsgerning får mange til uden videre at bruge forklaringsgraden til at konkludere, om der er tale om et godt lineært fit eller ej. Det skal man imidlertid være varsom med, som vi snart skal se. For det første skal forklaringsgraden som oftest være *meget* tæt på 1, for at punkterne bare ligger nogenlunde på linje. Et andet problem er, at forklaringsgraden ikke kan afsløre *systematiske afvigelser* fra en lineær sammenhæng! Derfor opfordres man altid til at lave en visuel inspektion af datapunkterne og den tilhørende regressionslinje og/eller lave et *residualplot*, som vi skal se hvad er nedenfor. Det ville vel også være underligt, hvis et enkelt tal skulle kunne indeholde al information om regressionen!

Definition 5.6 (Residual)

Residualet r_i for det *i*'te datapunkt (x_i, y_i) defineres som forskellen mellem punktets *y*-værdi y_i og den værdi $f(x_i) = a \cdot x_i + b$, som regressionslinjen forudsiger. Vi har med andre ord:

(8) $$r_i = y_i - f(x_i) = y_i - (a \cdot x_i + b)$$

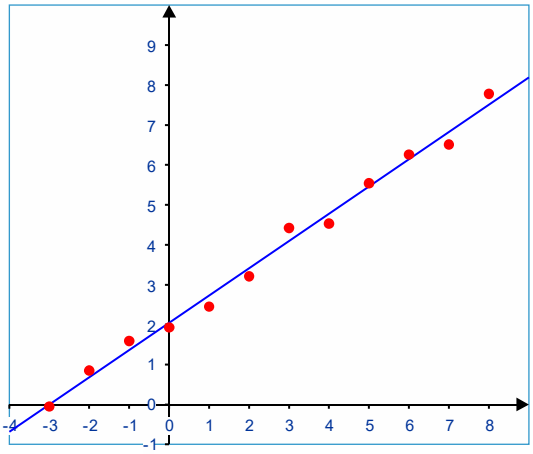


Rent grafisk er residualet r_i den lodrette afstand fra det *i*'te datapunkt (x_i, y_i) til regressionslinjen, regnet med fortegn. Hvis punktet ligger over linjen, er residualet positivt. Ligger punktet derimod under linjen, er residualet negativt. På figuren herover til venstre er de lodrette afstande markeret med en tynd grøn linje. På figuren til højre er residualet afbildet for hvert datapunkt. Det giver anledning til et såkaldt *residualplot* (bortset fra, at de grønne linjer normalt udelades). Det er nemmere at se punkternes variation i forhold til regressionslinjen, når man anvender residualplottet fremfor bare at kigge på det oprindelige plot. Vi skal betragte et eksempel, som illustrerer pointerne.

Eksempel 5.7

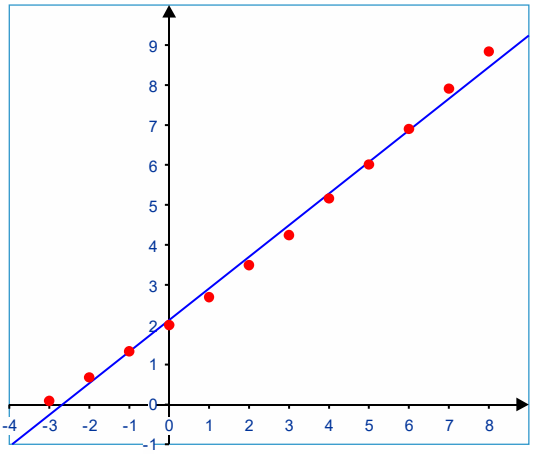
På de to øverste delfigurerne nedenfor ses to lineære regressioner med tilhørende residualplot umiddelbart under. Umiddelbart kan det øverst til højre synes som værende det bedste fit. Her er forklaringsgraden da også en smule tættere på 1 end tilfældet er på figuren øverst til venstre. MEN, når man ser bedre efter, så er der en vis systematik i, hvordan punkterne ligger i forholdet til regressionslinjen på figuren til højre. For lave og høje x-værdier ligger punkterne *over* linjen, mens de i midten ligger *under* linjen! Tendensen ses endnu tydeligere på residualplottet nedenfor: Ude til siden er residualerne *positive*, mens de i midten er *negative*. De systematiske afvigelser må give anledning til at betvivle, om der overhovedet er tale om en lineær sammenhæng. Eller måske skal der flere målinger til for at afgøre det. I situationen til venstre fordeler punkterne sig lidt "tilfældigt" omkring regressionslinjen, og residualplottet nedenfor fremstår kaotisk. Det er et godt tegn! Det giver større anledning til at tro, at der er tale om en lineær sammenhæng her.

Regression uden systematiske afvigelser



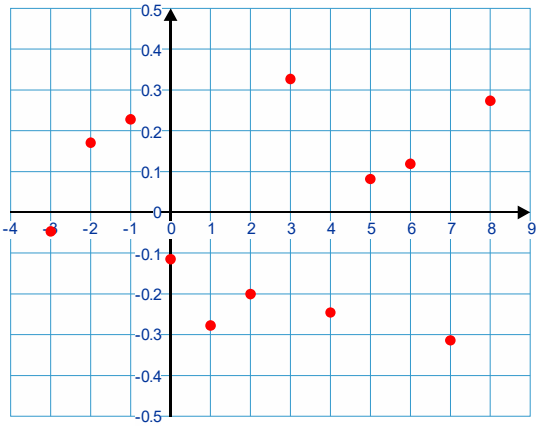
$R^2 = 0.9914$

Regression med systematiske afvigelser

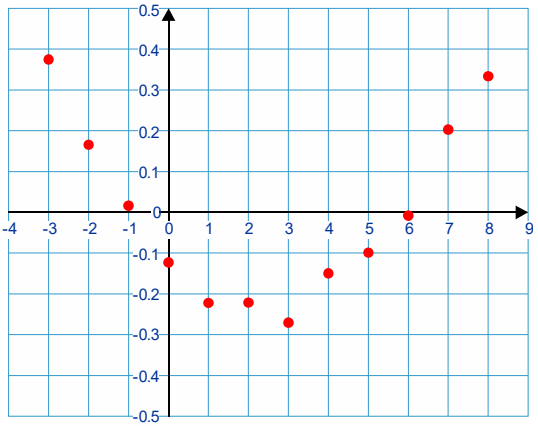


$R^2 = 0.9941$

Kaotisk residualplot uden systematik



Residualplot med systematik



I øvrigt vil gennemsnittet af residualerne altid give 0, som vist i sætning 5.9 lidt længere fremme. Det skal desuden nævnes, at man undertiden opdager *outliers* i data, altså datapunkter, som ligger usædvanligt yderligt. Det kan være rigtige data eller det kan være fejlmålinger. I sidstnævnte tilfælde bør man overveje at fjerne dem før regression.

Eksempel 5.8

Længden af en stængel for en hurtigvoksende plante blev målt hver anden dag i et par uger. Det gav følgende resultat:

Dage	1	3	5	7	9	11	13	15
Højde (cm)	43,5	49,1	51,9	56,8	62,7	66,0	72,3	74,7

Normalt vil man bruge sit CAS-værktøj til at finde a og b for regressionslinjen. For denne ene gang vil vi dog demonstrere brugen af formlerne (3), (4) og (5):

$$\overline{x} = \frac{1}{n}\cdot\sum_{i=1}^{n} x_i = \frac{1+3+5+7+9+11+13+15}{8} = 8$$

$$\overline{y} = \frac{1}{n}\cdot\sum_{i=1}^{n} y_i = \frac{43{,}5+49{,}1+51{,}9+56{,}8+62{,}7+66{,}0+72{,}3+74{,}7}{8} = 59{,}625$$

$$\overline{x\cdot y} = \frac{1}{n}\sum_{i=1}^{n} x_i\cdot y_i = \frac{1\cdot 43{,}5+3\cdot 49{,}1+5\cdot 51{,}9+7\cdot 56{,}8+9\cdot 62{,}7+11\cdot 66{,}0+13\cdot 72{,}3+15\cdot 74{,}7}{8}$$

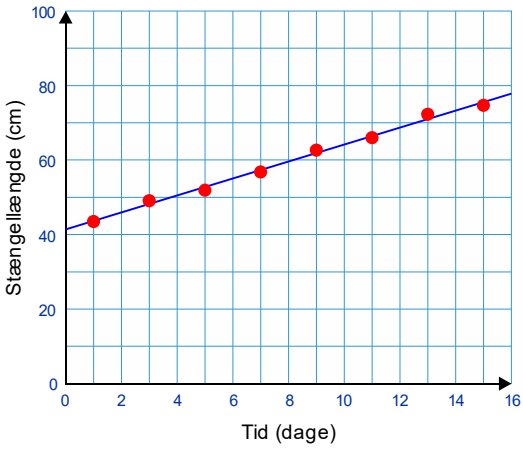
$$= 524{,}825$$

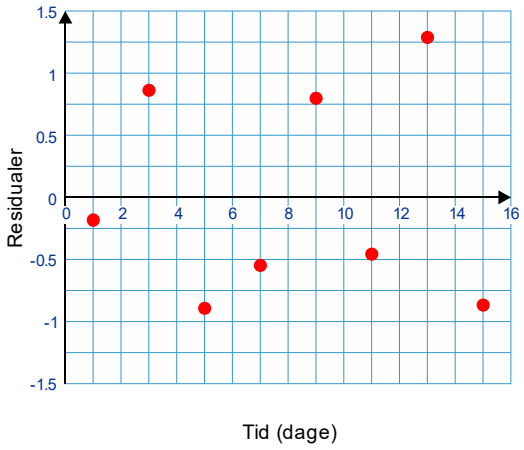
$$\overline{x^2} = \frac{1}{n}\cdot\sum_{i=1}^{n} x_i^2 = \frac{1^2+3^2+5^2+7^2+9^2+11^2+13^2+15^2}{8} = 85$$

$$a = \frac{\overline{x\cdot y}-\overline{x}\cdot\overline{y}}{\overline{x^2}-\overline{x}^2} = \frac{524{,}825-8\cdot 59{,}625}{85-8^2} = 2{,}2774$$

$$b = \overline{y}-a\cdot\overline{x} = 59{,}625-2{,}2774\cdot 8 = 41{,}4060$$

hvorved vi har $f(x) = 2{,}2774\cdot x + 41{,}4060$, hvis graf er regressionslinjen.





Som eksempel kan vi udregne residualet for 2. datapunkt, dvs. det i $x = 3$:

$$r_2 = y_2 - f(x_2) = y_2 - (a\cdot x_2 + b) = 49{,}1-(2{,}2774\cdot 3+41{,}4060) = 0{,}8619$$

På den venstre del af figuren på forrige side kan vi se grafen for datapunkterne plus regressionslinje. For det første ser vi, at punkterne ligger tæt på regressionslinjen, men hvad der faktisk er vigtigst for, at vi kan afgøre, om der er tale om en lineær sammenhæng er, at punkterne ligger lidt tilfældigt under og over linjen. Der er ikke områder af x-aksen, hvor punkterne gennemgående ligger på den ene eller den anden side. Der synes ikke at være *systematiske afvigelser*. Det afspejler sig også i residualplottet i den højre delfigur på forrige side. Her ligger punkterne rimeligt kaotisk og usystematisk henholdsvis over og under x-aksen. Det taler umiddelbart for en lineær sammenhæng.

Lad os afslutningsvist udregne forklaringsgraden med formlen (7). Før vi kan gøre det, mangler vi dog lige at udregne en enkelt kvadratsum:

$$\overline{y^2} = \frac{43{,}5^2+49{,}1^2+51{,}9^2+56{,}8^2+62{,}7^2+66{,}0^2+72{,}3^2+74{,}7^2}{8} = 3664{,}70$$

$$R^2 = a^2\cdot\frac{\overline{x^2}-\overline{x}^2}{\overline{y^2}-\overline{y}^2} = 2{,}2774^2\cdot\frac{85-8^2}{3664{,}70-59{,}625^2} = 0{,}9941$$

Normalt vil man dog lade ens regneværktøj regne det ud. Vi ser, at forklaringsgraden er tæt på 1, hvilket er et tegn på, at punkterne ligger tæt på regressionslinjen. Men størrelsen kan altså ikke bruges til at afsløre systematiske afvigelser fra en lineær sammenhæng. Derfor er det normalt bedre at se på residualplottet.

□

Sætning 5.9

Middelværdien af alle residualerne ved en lineær regression er 0, dvs. $\sum_{i=1}^{n} r_i = 0$.

Hjælp: Det er lidt teknisk. Undervejs sætter vi konstante faktorer udenfor summerne:

$$\sum_{n=1}^{n} r_i = \sum_{n=1}^{n}\left(y_i - f(x_i)\right) = \sum_{n=1}^{n}\left(y_i - a\cdot x_i - b\right) = \sum_{n=1}^{n} y_i - a\cdot\sum_{n=1}^{n} x_i - b\cdot\sum_{i=1}^{n} 1$$
$$= n\cdot\overline{y} - a\cdot n\cdot\overline{x} - b\cdot n = n\cdot\overline{y} - a\cdot n\cdot\overline{x} - \left(\overline{y} - a\cdot\overline{x}\right)\cdot n = 0$$

hvor vi desuden har benyttet definitionerne (5) samt egenskaben (3) for regressionslinjen.

□

5.3 Den simple lineære regressionsmodel

Vi skal nu betragte en udbygning af regressionsmodellen fra afsnit 5.2, hvor der blot var tale om en række datapunkter og ikke andet. I den simple lineære regressionsmodel vil vi antage, at der er lidt mere struktur.

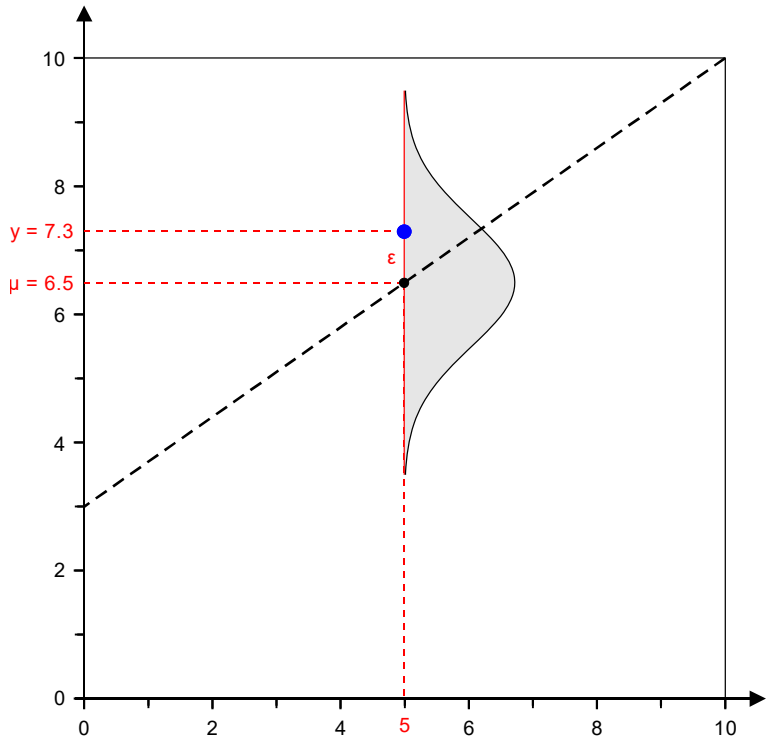
Definition 5.10 (Den simple lineære regressionsmodel)

Den *simple lineære regressionsmodel* er formelt set en sammenhæng mellem en forklarende variabel x og en stokastisk responsvariabel Y på formen:

(9) $$Y = \alpha \cdot x + \beta + \varepsilon$$

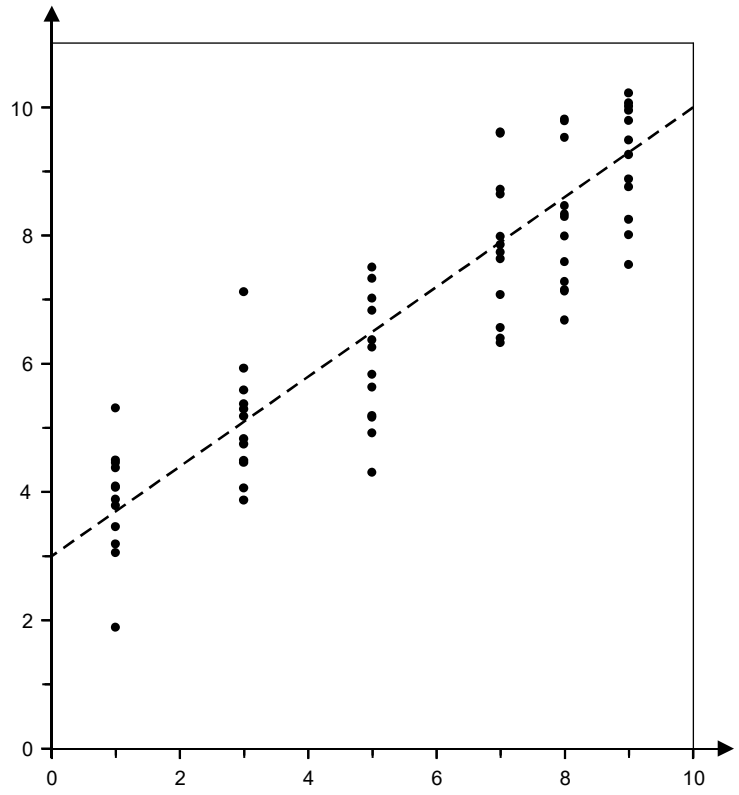
hvor det sidste led ε er en normalfordelt stokastisk variabel med middelværdi 0 og spredning σ, altså $\varepsilon \sim N(0, \sigma)$. Man kan tænke på leddet, som en form for "støj".

En værdi y af Y afhænger udover af x også af værdien af den stokastiske variabel ε. Lad os kigge på eksemplet $Y = 0{,}7 \cdot x + 3 + \varepsilon$, og antage at spredningen σ er 1. Antag desuden, at $x = 5$. Der er mange muligheder for hvilken værdi ε kan antage. Vi kan for eksempel forsøge os med på tilfældig måde at udtrække en værdi fra normalfordelingen med middelværdi 0 og spredning $\sigma = 1$. Lad os sige, at det resulterede i en værdi på 0,8 for ε. Det ville give følgende y-værdi: $y = 0{,}7 \cdot 5 + 3 + 0{,}8 = 7{,}3$, hvorved vi har det blå punkt på figuren med koordinater $(5; 7{,}3)$. Den lodrette afstand fra det blå punkt til linjen med ligning $y = \alpha \cdot x + \beta$ er altså støjen på 0,8.



"Klokkekurven" for normalfordelingen med middelværdi 0 og spredning 1 er indtegnet for at indikere, at det er mere sandsynligt at udtrække støjled, som er tæt på 0, end at udtrække støjled, som ligger langt fra 0. Det kan vi også indse ved ikke blot at generere én mulig støjværdi, men mange tilfældige støjværdier. Det giver anledning til en række mulige y-værdier for hver x-værdi, som vist på figuren på næste side. Simuleringerne er endda udført for seks forskellige x-værdier. I øvrigt kan vi nemt se, at Y i punktet x er en normalfordelt stokastisk variabel med middelværdi $\mu = \alpha \cdot x + \beta$ og med spredning σ. Med andre ord: $Y \sim N(\alpha \cdot x + \beta, \sigma)$.

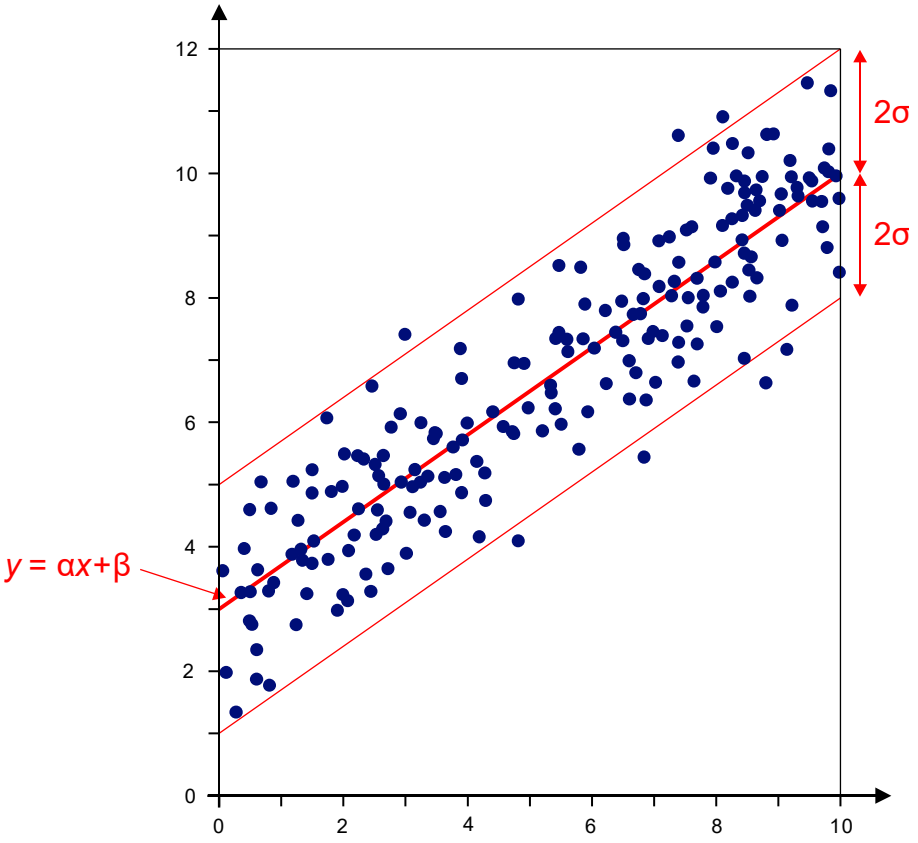
I de konkrete x-værdier i eksemplet er Y i punktet x altså en normalfordelt stokastisk variabel med middelværdi $\mu = 0{,}7 \cdot x + 3$ og spredning 1.



Vi ser, at den simple lineære regressionsmodel har en ekstra struktur i forhold til situationen i afsnit 5.2, nemlig kravet om den normalfordelte støj! Det er desuden vigtigt, at bemærke, at spredningen σ for støjleddet ε er er den samme for alle x-værdier. Lad os for eksempel antage, at vi vælger en masse x-værdier (her 200 styk) og for hver af dem genererer en tilfældig y-værdi ved hjælp af den tilhørende normalfordeling. Derved fås en hel "sky" af punkter, som fordeler sig om linjen $y = 0{,}7 \cdot x + 3$. Det kunne for eksempel give anledning til plottet øverst på næste side. Spredningen er stadig 1. Ifølge egenskaber for normalfordelingen vil ca. 95% af observationerne ligge indenfor 2σ fra middelværdien. Det passer fint med, at hele 193 ud af de 200 punkter ligger mellem de to røde spredningslinjer, som er anbragt i den lodrette afstand 2σ fra den rigtige regressionslinje. Det svarer til 96,5%. Altså fin overensstemmelse. Det er klart, at man ikke vil få den samme procent hver gang, da der er tale om tilfældigheder, men har man som her mange punkter, vil *de store tals lov* betyde, at resultatet ofte vil ligge meget tæt på 95%.

Vi indser altså, at den simple lineære regressionsmodel giver anledning til noget mere struktur end den fra afsnit 5.2. Her var der blot tale om punkter uden krav til, hvor de kom fra. Der er større krav til at data skal adlyde den simple lineære regressionsmodel. Til gengæld får man også mere ud af modellen. Senere skal vi blandt andet se, hvordan man ud fra data kan opstille konfidensintervaller for hældningskoefficienten for den "rigtige" regressionslinje.

Til sidst skal det lige nævnes, at man sagtens kan have flere punkter med samme x-værdi. Det giver modellen mulighed for. Det kunne være data fra samfundsvidenskaberne, hvor man ofte vil se datapunkter, med den samme 1. koordinat. De tilhørende 2. koordinater kan sagtens være forskellige her.



Jo større σ er, jo mere spredt vil datapunkterne almindeligvis ligge i forhold til den "rigtige" regressionslinje $y = \alpha \cdot x + \beta$, regnet i lodret afstand. Det er desuden vigtigt at bemærke, at mens middelværdierne normalt varierer, så skal spredningen σ være den samme i alle x-værdier. Det ligger i selve modellens definition. Vi er nu klar til at definere problemstillingen med regression i tilfældet med den simple lineære regressionsmodel.

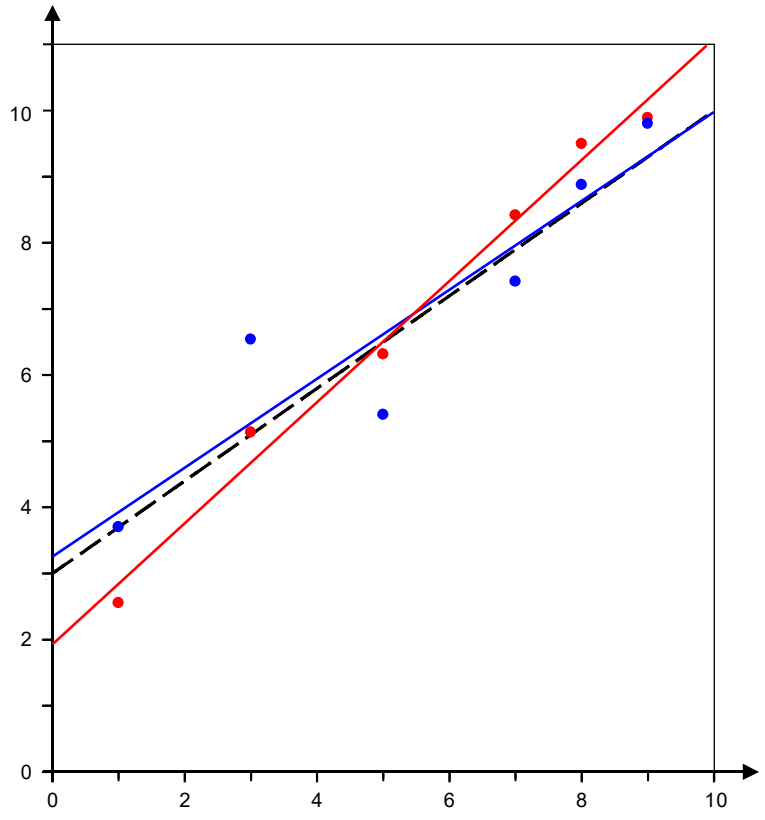
Definition 5.11 (Et regressionsproblem)

Normalt er den lineære komponent $y = \alpha \cdot x + \beta$ fra den simple lineære regressionsmodel ukendt, ligesom spredningen σ også normalt er det. Antag, at man har givet en række datapunkter $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, hvor y_i'erne kan betragtes som værdier af den stokastiske variabel Y i de tilhørende x-værdier $x_1, x_2, \ldots, x_n$. Det handler om at udnytte datapunkterne til at *estimere* de ukendte parametre α og β og eventuelt også estimere σ. Man kan eventuelt betegne linjen $y = \alpha \cdot x + \beta$ som den "rigtige" regressionslinje.

For at forstå problematikken yderligere, arbejder vi videre med den simple lineære regressionsmodel givet ved den stokastiske variabel $Y = 0{,}7 \cdot x + 3 + \varepsilon$, hvor støjen er normalfordelt med middelværdi 0 og spredning $\sigma = 1$. Vi betragter seks x-værdier: 1, 3, 5, 7, 8 og 9. Middelværdierne for Y i disse seks punkter beregner vi i tabellen nedenfor.

x_i	1	3	5	7	8	9
$\mu_i = 0{,}7 \cdot x_i + 3$	3,7	5,1	6,5	7,9	8,6	9,3

Vi tænker os nu genereret to sæt af tilhørende y-værdier ud fra normalfordelingen. Det kunne for eksempel give det resultat, som man ser plottet på figuren herunder: Et sæt af røde punkter og et sæt af blå punkter. For hvert sæt af punkter bestemmer vi den tilhørende regressionslinje - ligesom i afsnit 5.2. Regressionslinjerne er ligeledes farvet henholdsvis rød og blå, så man kan se, hvilket datasæt de hører til. Hvad vi observerer er, at ingen af regressionslinjerne falder sammen med den "rigtige" regressionslinje med forskriften $y = 0{,}7 \cdot x + 3$. Den blå linje er dog tæt på. Det illustrerer, at den estimerede regressionslinje afhænger af datasættet med de "tilfældigt udvalgte" y-værdier.



Vi kan altså nemt finde et estimat for den "rigtige" regressionslinjelinje $y = 0{,}7 \cdot x + 3$, men hvad der er vigtigere: På grund af den ekstra struktur, bliver vi også i stand til at give et estimat for såvel spredningen σ som et konfidensinterval for hældningskoefficienten. Det ser vi på i de næste par afsnit.

Bemærkning 5.12

Det skal nævnes, at regressionslinjen egentligt fremkommer ved en proces, som betegnes *Maximum Likelihood Estimation*, men da den metode munder ud i nøjagtigt den samme regressionslinje, som metoden fra afsnit 5.2 giver, så gør vi ikke mere ud af det her.

□

Bemærkning 5.13

Bemærk, at en værdi for støjen ε *ikke* skal forveksles med et residual *r*. Sidstnævnte er forskellen mellem et datapunkts *y*-koordinat og den *y*-værdi, som den *estimerede regressionslinje* $y = a \cdot x + b$ forudsiger, mens værdien af ε er forskellen mellem datapunktets *y*-koordinat og den *y*-værdi, som den "rigtige" regressionslinje $y = \alpha \cdot x + \beta$ forudsiger. Der er altså to forskellige linjer involveret!

□

Bemærkning 5.14

Det bør lige præciseres, at det i definition 5.10 er underforstået, at støjen ε i en *x*-værdi er *uafhængig* af støjen i en anden *x*-værdi.

□

5.4 Residualspredning

I det følgende vil vi vende os mod den normale problemstilling, hvor såvel den lineære funktion $f(x) = \alpha \cdot x + \beta$ som spredningen σ er ukendt. Kun de *n* datapunkter kendes. Der foretages lineær regression på datapunkterne. Herved fås regressionslinjen $y = a \cdot x + b$, som er et estimat for linjen $y = \alpha \cdot x + \beta$. Da vi ikke kender den "rigtige" linje, så bruger vi det næstbedste vi har, nemlig den regressionslinje, som datapunkterne giver os.

Definition 5.15 (Residualspredning)

I den simple lineære regressionsmodel defineres *residualspredningen s* for et datasæt $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ således:

(10)
$$s = \sqrt{\frac{r_1^2 + r_2^2 + \ldots + r_n^2}{n-2}}$$

hvor $r_i = y_i - (a \cdot x_i + b)$ er residualerne for datapunkterne i forhold til den estimerede regressionslinje $y = a \cdot x + b$.

Residualspredningen *s* er et skøn eller estimat for spredningen σ i den *simple lineære regressionsmodel*. Grunden til, at man dividerer med $n-2$ og ikke *n*, ligger uden for denne notes opgave at forklare. En løs indikation er, at man mister to frihedsgrader, da punkterne indirekte også skal benyttes til at estimere α og β.

Eksempel 5.16 (Quiz)

En tilfældig udvalgt gruppe på 200 personer i alderen fra ca. 9 år til godt 30 år deltager i en quiz, hvor der stilles en række almene spørgsmål. Der er i alt 100 spørgsmål, så der kan samlet set tildeles point fra 0 til 100. En forsker påstår, at der er en lineær sammenhæng mellem alder og pointtal i quizzen. Det skal vi studere nærmere.

Alder (år)	23,5	11,9	...	16,1	19,0	17,3
Points	65	49		68	64	55

Dele af data fra filen alder_points.xlsx

I det følgende importeres data fra Excelfilen med navnet *alder_points.xlsx* i et CAS-værktøj med henblik på at foretage nogle analyser af data.

a) Foretag lineær regression på data. Angiv i den forbindelse værdier for a og b for regressionslinjen $y = a \cdot x + b$.

b) Lav et residualplot og bestem en værdi for residualspredningen. Hvor stor en procentdel af observationerne er *normale*, dvs. ligger inden for to residualspredninger fra det, som regressionslinjen forudsiger?

c) Vurder, om den lineære model ser ud til at kunne bruges til at beskrive data.

d) Undersøg hvorvidt residualerne er normalfordelte med henblik på at afgøre, om vi har at gøre med en simpel lineær regressionsmodel.

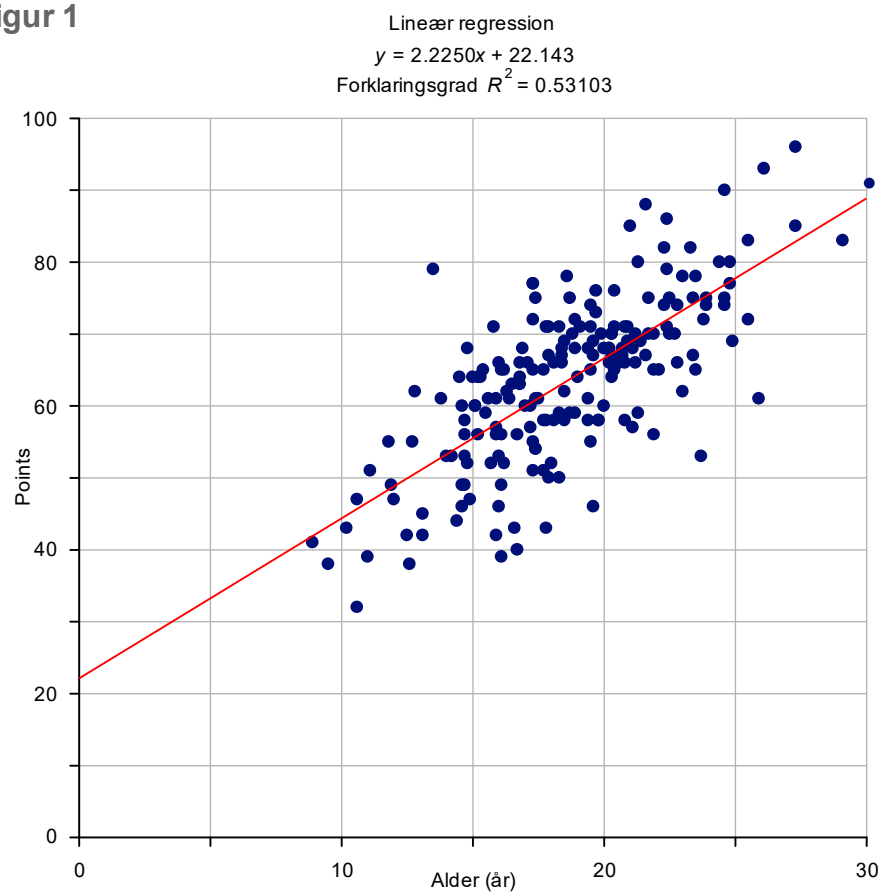
e) Er der nogen exceptionelle observationer?

Løsning: Vi skal overvejende anvende automatiserede værktøjer fra vores regneværktøj til at løse opgaven. Det antages, at læseren i forvejen er bekendt med manuel udregning af for eksempel residualer, som blev foretaget i det tidligere eksempel 5.8.

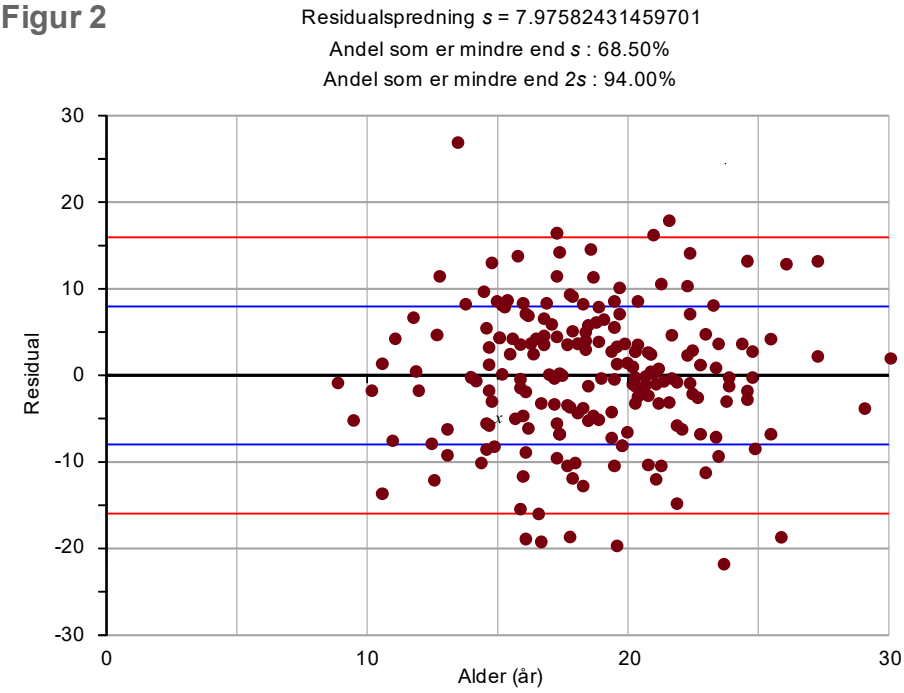
a) På figur 1 på næste side har vi udført lineær regression på data. Plottet viser data ligge som en oval sky. Vi har desuden, at $a = 2{,}2250$ og $b = 22{,}143$.

b) På figur 2 på næste side har vores regneværktøj afbildet et residualplot. Residualspredningen står angivet til at være ca. $s = 7{,}98$ (points). Da filen med data er lang, handler det om at finde en automatisk måde at optælle, hvor mange af punkterne, som har et residual i intervallet $[-2s, 2s]$. Værktøjet *plotResidualer* har faktisk allerede gjort arbejdet for os. Teksten over figur 2 viser, at 94.0% af observationerne er "normale". Der er endda tegnet to røde vandrette linjer, som afgrænser de observationer, som er normale. Faktisk er der også tegnet blå vandrette linjer, som afgrænser de observationer, som numerisk set har et residual på højst s. Sammenlagt 68,5% af observationerne ses at ligge i dette område.

Figur 1

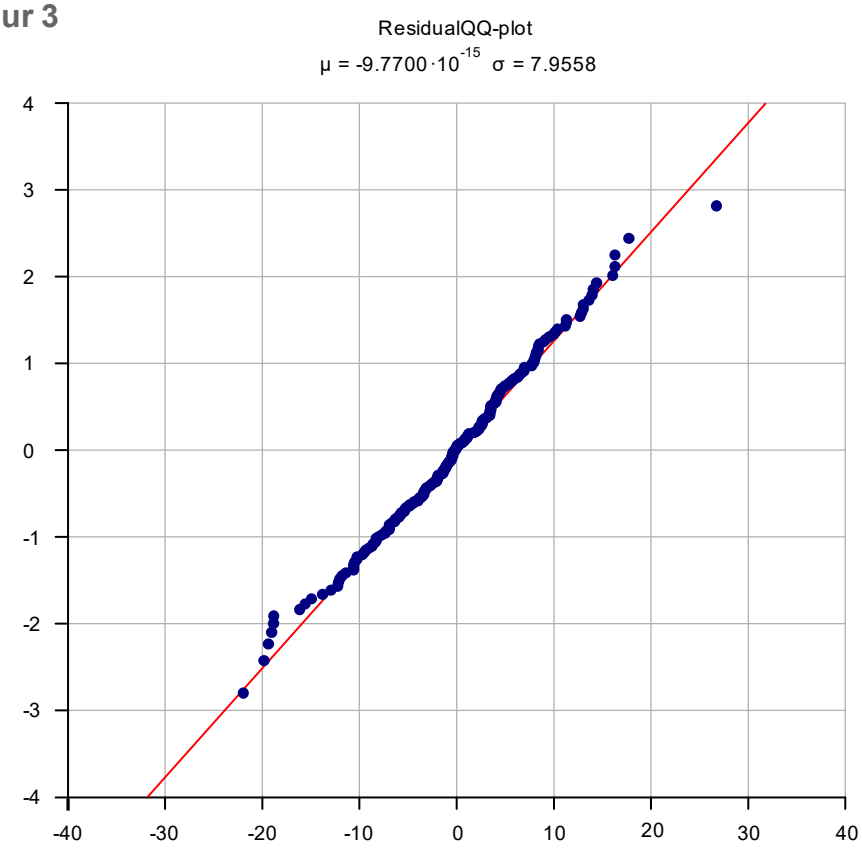


Figur 2



c) En visuel inspektion af residualernes fordeling fortæller, at punkterne ligger temmelig "tilfældigt" over/under x-aksen. Residualspredningen er ikke helt lille i forhold til modellens y-værdier og variationen i y-værdierne, men heller ikke for stor. Sammen med b) får det os til at konkludere, at den lineære model er anvendelig.

Figur 3



d) Dette spørgsmål er egentligt mest et spor til A-niveau. En ting er, at residualerne er "kaotisk" fordelte, men er de også normalfordelte? For at afgøre det er det hensigtsmæssigt at lave et såkaldt QQ-plot af residualerne. Nogle værktøjer har endda indbygget en facilitet, som klarer begge ting på en gang, dvs. udregner residualerne og dernæst plotter dem i et QQ-plot. Det kombinerede værktøj kunne for eksempel hedde *residualQQ-plot*. På figur 3 ovenfor er et sådant plot vist. Eftersom punkterne ca. ligger på en ret linje, konkluderer vi, at residualerne omtrent er normalfordelte.

e) En exceptionel observation er en, hvis residual numerisk set er større end $3s$. Figur 2 har ikke nogen indikation af det, så man må finde en metode at tælle dem op på. Det viser sig, at der er en enkelt observation, som er exceptionel. En person på kun 13,5 år har fået 79 points i quizzen, hvilket er knap 27 flere points end forventet i forhold til den lineære model.

□

Bemærkning 5.17 (Videnskabsteoretisk)

Det er vigtigt, at man overvejer om forudsætningerne for at bruge den simple lineære regressionsmodel overhovedet er opfyldt, før man begynder at bruge teorien for den. Ellers kan det ende med, at man når frem til noget meningsløst! For det første skal man have formuleret en *generel* lineær sammenhæng mellem to størrelser – foruden støjled. Dernæst skal man have en *stikprøve* i form af et datasæt. Stikprøven skal kunne siges eller antages at have være udtaget på *simpel tilfældig vis* uden *bias* (skævhed) fra populationen. Desuden skal man gerne ud fra konteksten i den konkrete situation kunne sandsynliggøre, at der er tale om uafhængig normalfordelt "støj" med konstant spredning σ, som er betin-

gelsen i den simple lineære regressionsmodel. Om residualerne virkelig er normalfordelte med konstant spredning, kan man danne sig et omtrentligt billede af ved at kigge på residualplottet, eller endnu bedre ved at lave et residualQQ-plot, som omtalt i eksempel 5.16. I eksempel 5.16 kunne man redegøre for kravet om en simpel tilfældig stikprøve ved at nævne, at de 200 deltagere var udvalgt tilfældigt og uden bias fra populationen.

Man bør ellers til stadighed have for øje, at vi her befinder os i statistikken, så vi argumenterer induktivt fra det *specielle* (en stikprøve) til det *generelle* ("populationen"), jf. afsnit 4.1. En anden ting, man skal huske er, at en eventuel lineær sammenhæng på ingen måde behøver være en *kausal* sammenhæng! Til slut skal det nævnes, at man normalt bør have en eller anden idé om, at der kan være en plausibel/mulig sammenhæng mellem de to betragtede størrelser. Ellers kan det ende med, at man registrerer en meningsløs, *spuriøs* sammenhæng.

Kendetegnende for en god lineær model er praktisk set:

- Residualspredningen *s* er ikke for stor.
- Residualerne er tilfældigt spredt over og under *x*-aksen.
- Langt hovedparten af observationerne har residualer i intervallet [–2s, 2s].
- Undertiden vil det også være relevant at have opfyldt, at residualspredningen ikke er for stor sammenlignet med modellens *y*-værdier og variation i *y*-værdier i det betragtede *x*-område. Er der måleusikkerhed på data, bør *s* også være lille i forhold til usikkerheden.

5.5 Ekstra: Konfidensinterval for hældning

Vi slutter kapitlet af med en smule ekstra, som er pensum i matematik på A-niveau STX. Det handler om at bestemme et konfidensinterval for den ukendte hældningskoefficient α for den lineære sammenhæng $y = \alpha \cdot x + \beta$ i den simple lineære regressionsmodel. Vi skal også se, hvordan man kan afgøre, om der overhovedet er nogen sammenhæng mellem de to variable x og y. Hvis hældningen er 0, vil vi normalt sige, at der ikke kan påvises nogen sammenhæng mellem de to variable. Til slut kigger vi lidt på *prædiktion*.

I den simple lineære regressionsmodel kender vi ikke parametrene α og β, og for den sags skyld heller ikke spredningen σ. Datapunkterne gav os et estimat for koefficienterne a og b i den estimerede regressionslinje $y = a \cdot x + b$. Spørgsmålet er, hvor nøjagtige, de mon er? Det viser sig, at vi kan angive et 95% konfidensinterval for blandt andet den ukendte hældning α. Problematikken med et konfidensinterval studerede vi allerede i afsnit 4.5. I denne situation tager den følgende form: En *stikprøve* svarer i denne situation til givet en række x-værdier $x_1, x_2, \ldots, x_n$ at udtrække en række tilhørende y-værdier $y_1, y_2, \ldots, y_n$ med tilfældig, normalfordelt støj, efter retningslinjerne i definition 5.10. Et 95% konfidensinterval har i den forbindelse følgende egenskab: Hvis man foretager i tusindvis af stikprøver og for hver stikprøve beregner et tilhørende konfidensinterval, så vil omkring 95% af disse intervaller i praksis indeholde den korrekte hældning α. Det er lidt abstrakt.

Et udtryk for konfidensintervallet er angivet i sætningen nedenfor for fuldstændighedens skyld. Det vil dog være alt for kompliceret at bevise eller gøre nøjere rede for udtrykket. Intervallet indeholder desuden en fraktil fra den såkaldte *t*-fordeling, som ikke er pensum i gymnasiet. Når vi på A-niveau regner opgaver med konfidensintervaller, vil vi derfor ikke benytte formlen i (11). I stedet vil vi bruge indbyggede faciliteter i vores CAS-værktøj. Bemærk dog, at det er vigtigt at forstå, hvad et konfidensinterval i det hele taget er for noget, som beskrevet ovenfor!

Sætning 5.18 (95% konfidensinterval for hældningen)

Givet et datasæt $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ fra en simple lineær regressionsmodel. Et 95% konfidensinterval for den korrekte hældning α er givet ved:

(11) $$[a-w, a+w] \quad \text{hvor} \quad w = t_{0.975,n-2} \cdot \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}}$$

hvor a er hældningen for regressionslinjen hørende til de n datapunkter, s er residualspredningen og $t_{0.975.n-2}$ er 0,975-fraktilen for *t*-fordelingen med $n-2$ frihedsgrader.

Eksempel 5.19 (Konfidensinterval for hældningskoefficienten)

Vi vender tilbage til situationen i eksempel 5.16, hvor vi bestemte hældningskoefficienten for regressionslinjen for sættet af datapunkter til at være $a = 2{,}2250$. Nu ønsker vi at bestemme det til datasættet hørende 95% konfidensinterval for den rigtige hældningskoefficient α. I dette tilfælde benytter vi et værktøj, som hedder *testLin*. Det fodres med datasættet og giver det output, som ses i boksen herunder. Heri befinder der sig en masse information, hvoraf vi dog kun skal bruge en smule, nemlig de to værdier, som står ud for *Nedre 95.00%* og *Øvre 95.00%* i søjlen *a*. Det er nemlig endepunkterne i vores 95% konfidensinterval, som dermed bliver $[1{,}931966;\ 2{,}518034]$. Der er altså 95% sandsynlighed for at dette interval "rammer" vores rigtige, ukendte hældning α.

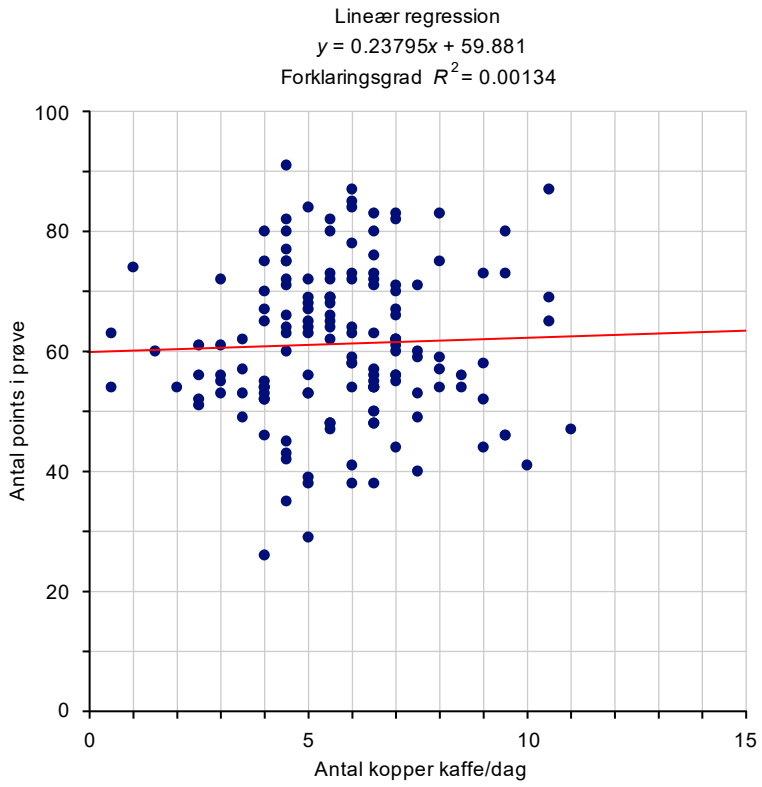
Data fra *testLin* værktøjet:

	a	b
Koefficient	2.225000	22.142637
Standardfejl	0.148596	2.828191
t-stat	14.973491	7.829258
p-værdi	0.000000	0.000000
Nedre 95.00%	1.931966	16.565395
Øvre 95.00%	2.518034	27.719879
Frihedsgrader	198	

Da intervallet udelukkende indeholder positive værdier, kan vi godtage et udsagn om, at alderen har en positiv betydning for, hvor mange points, man opnår i quizzen.

Eksempel 5.20 (Ingen sammenhæng)

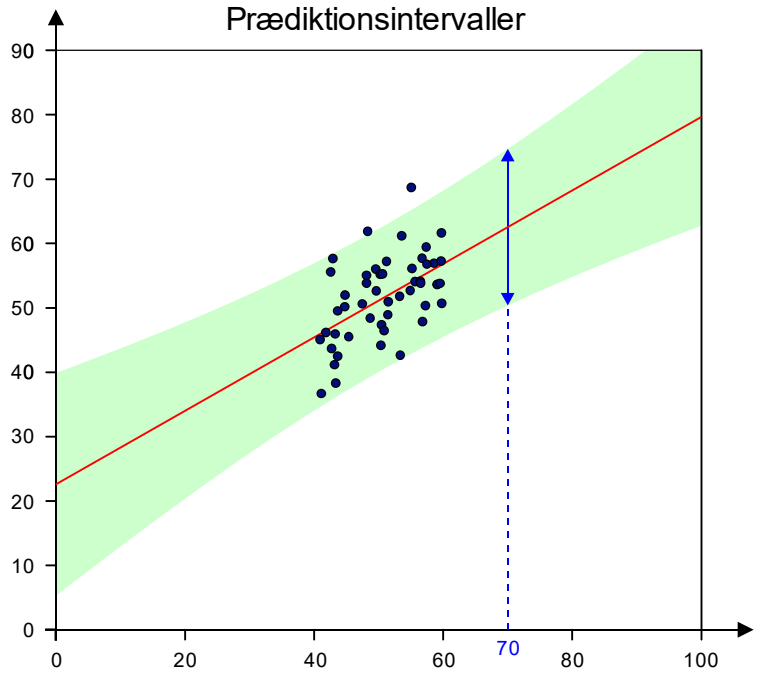
En kaffeproducent fremsætter den påstand, at der er en positiv lineær sammenhæng mellem, hvor mange kopper kaffe, man drikker, og hvor godt man klarer sig til en landsdækkende matematikeksamen. For at teste det, udtrækker man på tilfældig vis 150 elever ud af de flere tusinde, som skal til eksamen og som er kaffedrikkere. Hver person undersøger i en periode, hvor mange kopper kaffe denne gennemsnitligt drikker pr. dag. Antallet opgives med en nøjagtighed på 0,5 kop. Der gives fra 0 til 100 points i matematikprøven. Når prøveresultaterne foreligger, foretager man lineær regression på data for at afgøre, hvorvidt kaffeproducenten har ret. Plottet ses på figuren nedenfor. Man får en svagt positivt hældende linje. Men er den positive hældning signifikant? For at afgøre det, udregner vi et 95% konfidensinterval for den rigtige hældningskoefficient α. Resultatet af *testLin* værktøjet ses i boksen nedenfor. Vi får intervallet $[-0{,}815913; 1{,}291816]$. Eftersom 0 befinder sig i dette interval, vælger vi at sige, at der *ingen sammenhæng* er mellem kaffeindtaget og hvor mange points, man opnår til matematikeksamen. Det skal lige nævnes, at de øvrige plots: Residualplottet og residualQQ-plot begge taler for at betingelserne for den simple lineære regressionsmodel er opfyldt. Vi undlader at vise disse plots her.



	a	b
Koefficient	0.237951	59.880826
Standardfejl	0.533299	3.238375
t-stat	0.446187	18.491008
p-værdi	0.656114	0.000000
Nedre 95.00%	-0.815913	53.481400
Øvre 95.00%	1.291816	66.280253
Frihedsgrader	148	

I øvrigt skal det nævnes, at endepunkterne for konfidensintervallet for konstantleddet β står anført i søjlen *b*. Til sidst skal det lige nævnes, at det er karakteristisk for en situation uden sammenhæng, at skyen af datapunkter danner en vandretliggende oval eller cirkel, hvorimod skyen i tilfælde af en signifikant positiv eller signifikant negativ hældning vil danne en "hældende oval".

□



Prædiktion

Til slut skal det nævnes, at man ofte i praksis vil bruge den estimerede regressionslinje, som fås ud fra datapunkterne, til at give et bud på en mulig *y*-værdi, dvs. man indsætter *x*-værdien i $y = a \cdot x + b$. Man skal dog huske, at vi her benytter regressionslinjen og ikke den rigtige, ukendte lineære sammenhæng $y = \alpha \cdot x + \beta$. Dertil kommer også en mulig normalfordelt "støj", som jo er et indbygget element i den simple lineære regressionsmodel. Spørgsmålet er, hvor meget unøjagtigheden i regressionslinjen samt støjen betyder for den statistiske usikkerhed, når man skal *forudsige* en mulig *y*-værdi for *Y*. Det viser sig, at man også her kan angive et 95% konfidensinterval: et interval, som med ca. 95% sikkerhed "rammer" den værdi, som *Y* vil antage, inklusiv støj. Situationen er illustreret på figuren ovenfor. Her har vi 50 datapunkter, som har en vis spredning (residualspredningen er 5,44). Regressionslinjen hørende til datapunkterne er tegnet med rødt. Det grønne bagvedliggende område fortæller noget om nævnte konfidensinterval. Som eksempel er markeret konfidensintervallet for værdien $x = 70$. Hvordan disse konfidensintervaller udregnes, er for kompliceret til at blive gennemgået her. Man ser dog, at konfidensintervallets bredde er større i enderne af det afbildede område. Det skyldes, at regressionslinjen jo (sandsynligvis) ikke er nøjagtigt den "rigtige" lineære sammenhæng $y = \alpha \cdot x + \beta$. Regressionslinjen afhænger af datapunkterne, hvilket betyder, at der løst sagt vil "rokkes" ved linjen alt efter hvilke datapunkter, der forekommer. Det vil ikke overraskende betyde større udslag i enderne. Usikkerheden er mindst omkring middelværdien $x = \overline{x}$. Udover

usikkerheden på linjen, så er der også den indbyggede usikkerhed på punkterne, som jo ligger fordelt omkring den rigtige linje med en vis spredning σ.

I øvrigt bør man helst ikke *ekstrapolere* for meget udenfor serien af datapunkter. Er man for eksempel sikker på, om den simple lineære model også holder her? Vi vil stoppe diskussionen her. Man kan udvikle meget mere matematik indenfor emnet, hvis man vil.

Appendiks A. Population og stikprøve størrelser

I dette appendiks skal vi kigge på forskellige størrelser i forbindelse med både populationer og stikprøver: *middelværdi*, *varians* og *spredning*.

Ofte er vi interesseret i at bestemme størrelserne for en hel population, men må ofte på grund af manglende resurser nøjes med at give et estimat af populationsstørrelserne ud fra en stikprøve. Første betingelse for, at det kan gå godt er, at stikprøven er *tilfældig* og *repræsentativ*, dvs. ikke har nogen *skævhed*, også kaldet *bias*. Lad os antage, at populationen indeholder N elementer. Da har vi følgende definitioner:

Definition A1 (Størrelser for populationen)

Populationens *middelværdi* μ er defineret ved:

$$\mu = \frac{1}{N}\cdot\sum_{i=1}^{N} x_i = \frac{x_1+x_2+\ldots+x_N}{N} \tag{A1}$$

Populationens *varians* σ^2 er defineret ved:

$$\sigma^2 = \frac{1}{N}\cdot\sum_{i=1}^{N} (x_i-\mu)^2 = \frac{(x_1-\mu)^2+(x_2-\mu)^2+\ldots+(x_N-\mu)^2}{N} \tag{A2}$$

Populationens *spredning* σ er defineret ved:

$$\sigma = \sqrt{\frac{1}{N}\cdot\sum_{i=1}^{N} (x_i-\mu)^2} = \sqrt{\frac{(x_1-\mu)^2+(x_2-\mu)^2+\ldots+(x_N-\mu)^2}{N}} \tag{A3}$$

Vi har allerede defineret størrelserne i forbindelse med stokastiske variable i afsnit 2.4, og ovenstående er i pæn overensstemmelse med disse.

Nu til de tilsvarende værdier for en stikprøve: *Stikprøve-middelværdi*, *stikprøvevarians* og *stikprøvespredning* omtales også ofte som henholdsvis den *empiriske middelværdi*, den *empiriske varians* og den *empiriske spredning*.

Definition A2 (Stikprøve-middelværdi)

Stikprøve-middelværdien for stikprøven bestående af observationerne $x_1, x_2, \ldots, x_n$ er defineret ved:

$$\overline{x} = \frac{1}{n}\cdot\sum_{i=1}^{n} x_i = \frac{x_1+x_2+\ldots+x_n}{n} \tag{A4}$$

Vi ser, at stikprøve-middelværdien eller den empiriske middelværdi blot er *gennemsnittet* af observationerne i stikprøven!

Definition A3 (Stikprøvevarians og stikprøvespredning)

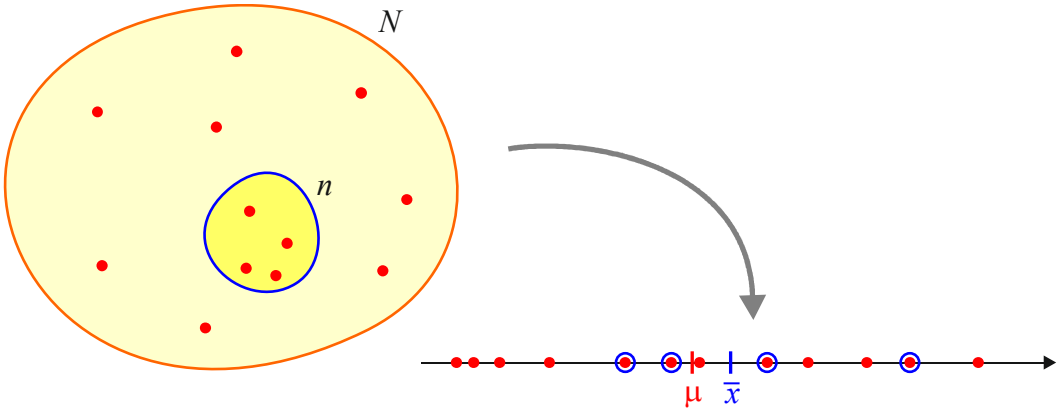
Stikprøvevariansen for stikprøven bestående af observationerne $x_1, x_2, \ldots, x_n$ er defineret ved:

$$(A5) \qquad s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n-1}$$

og *stikprøvespredningen* er defineret ved:

$$(A6) \qquad s = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2} = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n-1}}$$

Figuren herunder illustrerer problematikken mellem population og stikprøve: Man ønsker en værdi for populationsmiddelværdien μ, men da man kun undersøger en stikprøve, kan man kun få et estimat $\overline{x}$ for denne.



Bemærkning A4

Man kan måske undre sig over, at der i udtrykket for stikprøvevariansen og stikprøvespredningen figurerer et $n-1$ i brøkens nævner frem for bare et n. Lidt løst sagt kan det forklares ved, at hvis man i nævneren anvendte n, så ville stikprøvevariansen i gennemsnit over mange stikprøver skyde lidt for lavt i forhold til den *korrekte* varians σ^2. En mere stringent og kompliceret forklaring kan findes i mit tillæg på min hjemmeside (se [L1] i listen over links til hjemmesider bagerst i denne ebog).

Bemærkning A5

I afsnit 4.6 kan desuden studeres begrebet *konfidensinterval* i forbindelse med en middelværdi. Dette interval fortæller lidt om den statistiske usikkerhed på en stikprøve-middelværdi.

Eksempel A6

Gårdejer Jensen ejer en æbleplantage og ønsker at undersøge, hvordan vægten af æblerne fordeler sig i plantagen. Til det formål udtager han en stikprøve på 30 æbler spredt ud over forskellige træer på hans areal. Vægten af de udvalgte æbler er som følger (i gram):

145, 143, 167, 162, 171, 148, 143, 153, 184, 141, 141, 156, 149,
155, 153, 155, 172, 148, 150, 157, 174, 179, 160, 142, 145, 168,
152, 171, 139, 153

Først bestemmer vi stikprøve-middelværdien:

$$\overline{x} = \frac{1}{n}\cdot\sum_{i=1}^{n} x_i = \frac{145+143+\ldots+153}{30} = 155{,}867$$

Den gennemsnitlige vægt af æblerne i stikprøven er altså 155,9 gram og er det bedste bud, vi kan give på den rigtige gennemsnitsvægt μ af alle æblerne på hans plantage. For at få et estimat på den sande spredning (standardafvigelse) σ benytter vi stikprøvespredningen (også kaldt den *empiriske spredning*):

$$\begin{aligned} s &= \sqrt{\frac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i-\overline{x})^2} \\ &= \sqrt{\frac{(145-155{,}867)^2+(143-155{,}867)^2+\ldots+(153-155{,}867)^2}{30-1}} \\ &= 12{,}4 \end{aligned}$$

Stikprøvespredningen er altså 12,4 gram.

□

Bemærkning A7

Hvis æblernes vægt er *normalfordelt*, hvilket sikkert er en rimelig antagelse, kan man sige mere. Her får man nemlig (jf. opgave 327), at ca. 68% af æblerne vil have en vægt, som er indenfor 1 spredning fra middelværdien, og at ca. 95% af æblerne vil have en vægt, som er indenfor 2 spredninger fra middelværdien.

Øvelse A8

Landmand Olsen vil undersøge, hvordan gulerødderne fra hans mark fordeler sig, hvad angår længde. Det viser sig nemlig, at den betaling han kan modtage fra distributørerne, afhænger af, at gulerøddernes længde holder sig inden for et vist længdeinterval.

Han foretog en tilfældig stikprøve på 40 gulerødder fra forskellige steder på hans mark. Det gav følgende længder, underforstået i cm, og angivet med decimalpunktum:

21.7, 18.7, 18.9, 22.3, 22.2, 15.7, 17.3, 15.1, 14.1, 20.3, 19.5, 19.7, 22.3, 23.4, 21.5, 23.6, 15.9, 23.9, 19.6, 20.5, 16.4, 17.4, 15.6, 16.9, 17.7, 13.7, 20.0, 20.8, 20.4, 19.7, 21.6, 24.7, 18.4, 22.4, 18.7, 16.3, 18.8, 18.3, 19.1, 14.1

a) Hvad er stikprøven og populationen i denne sammenhæng?
b) Bestem stikprøve-middelværdien.
c) Bestem stikprøvespredningen.
d) Hvis du har lært om QQ-plot og normalfordelingen, kan du desuden foretage et QQ-plot med henblik på at vurdere, om gulerøddernes længde er normalfordelt.

Appendiks B. Forstå QQ-plot

Af og til får man brug for at afgøre, om noget givet datamateriale i form af en stikprøve er normalfordelt. Man kunne selvfølgelig tænke sig at gruppere data i intervaller og derefter se, om det tilhørende histogram kan fittes med en "klokkekurve", altså grafen for tæthedsfunktionen for normalfordelingen med middelværdi $\overline{x}$ og spredning s, jf. betegnelserne i Appendiks A. Det er imidlertid ikke altid nemt at afgøre, om et sådant fit er tilstrækkelig godt. Det samme ville være tilfældet, hvis man havde anvendt de kumulerede frekvenser og forsøgt at tilnærme sumkurven med grafen for fordelingsfunktionen for normalfordelingen. Det er nemmere at afgøre, om noget er omtrent *lineært*. Derfor anvender man ofte et såkaldt *QQ-plot*, eller som det også kaldes, et *fraktilplot*. Man ser undertiden også anvendt betegnelsen *probit-plot* på dansk. Vi holder os dog til betegnelsen QQ-plot. Q er en forkortelse for *Quantile* på engelsk. Det betyder *fraktil*.

Lad os kigge på et eksempel med henblik på at forstå, hvad et QQ-plot er. Vi tænker os foretaget en stikprøve med 10 elementer:

25,8 20,8 35,4 23,3 38,0 25,8 31,8 28,8 30,1 42,0

Vi kan eventuelt starte med at bestemme stikprøvens *empiriske middelværdi* $\overline{x}$ samt *stikprøvespredningen* s, som er omtalt i Appendiks A:

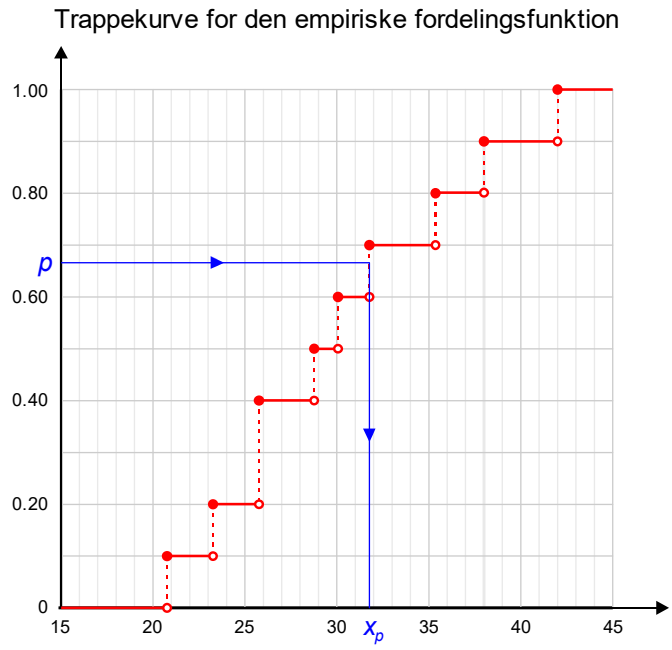
$$\begin{aligned}\overline{x} &= \frac{1}{n}\cdot\sum_{i=1}^{n} x_i \\ &= \frac{25{,}8+20{,}8+35{,}4+23{,}3+38{,}0+25{,}8+31{,}8+28{,}8+30{,}1+42{,}0}{10} \\ &= 30{,}180\end{aligned}$$

$$\begin{aligned}s &= \sqrt{\frac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i-\overline{x})^2} \\ &= \sqrt{\frac{(25{,}8-30{,}180)^2+(20{,}8-30{,}180)^2+(35{,}4-30{,}180)^2+\ldots+(42{,}0-30{,}180)^2}{10-1}} \\ &= 6{,}7193\end{aligned}$$

Men vi mangler at vurdere, om stikprøven med rimelighed kan tænkes at komme fra en normalfordeling. For at vurdere det, vil vi først tage fat i stikprøvens *empiriske fordelingsfunktion*. Den er defineret meget formelt i en ramme på næste side. Lad os se på stikprøven med de $n=10$ elementer ovenfor og se, hvordan fordelingsfunktionen ser ud i dette tilfælde. For ethvert x skal vi finde ud af, hvor mange af stikprøvens elementer eller observationer, som er mindre end eller lig med x. Derefter skal vi dividere med antal elementer, her 10. Så har vi $F_{emp}(x)$. For at klare dette er det hensigtsmæssigt at sætte elementerne i voksende rækkefølge:

20,8 23,3 25,8 25,8 28,8 30,1 31,8 35,4 38,0 42,0

Med denne sortering er det nemmere at overskue. Lad os se på $x = 33$. Vi ser hurtigt, at der i den sorterede række af elementer 7 elementer, som er mindre end eller lig med 33. Derfor haves: $F_{emp}(33) = \frac{7}{10}$. Sådan skal vi i princippet gøre for alle x-værdier. Man indser hurtigt, at det giver anledning til en *trappefunktion* (*stykvis lineær funktion*). Hver gang man kommer til en af de nye sorterede observationer, vil funktionsværdien vokse med et helt multiplum af $1/n$. Vi ser, at observationen 25,8 står anført to gange i den sorterede række. Derfor vil fordelingsfunktionen her vokse med $2/10 = 0{,}2$. Alle de øvrige observationer forekommer kun én gang, så her vil fordelingsfunktionen kun vokse med $1/10 = 0{,}1$. Grafen for fordelingsfunktionen bliver derfor en *trappekurve*:



Definition B1 (En stikprøves empiriske fordelingsfunktion)

Den *empiriske fordelingsfunktion* hørende til stikprøven bestående af observationerne $x_1, x_2, \ldots, x_n$ er defineret ved

(B1) $$F_{emp}(x) = \frac{\#\{i \mid x_i \le x\}}{n}$$

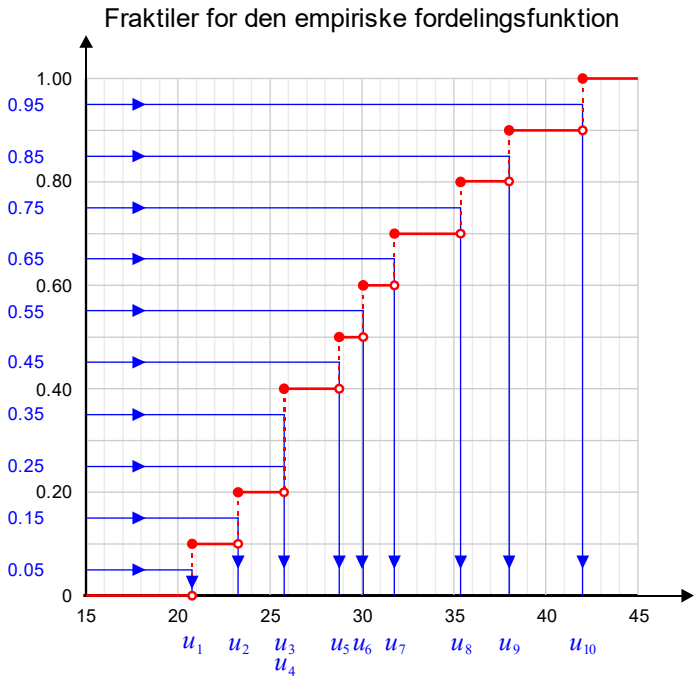
hvor # betyder "antallet af".

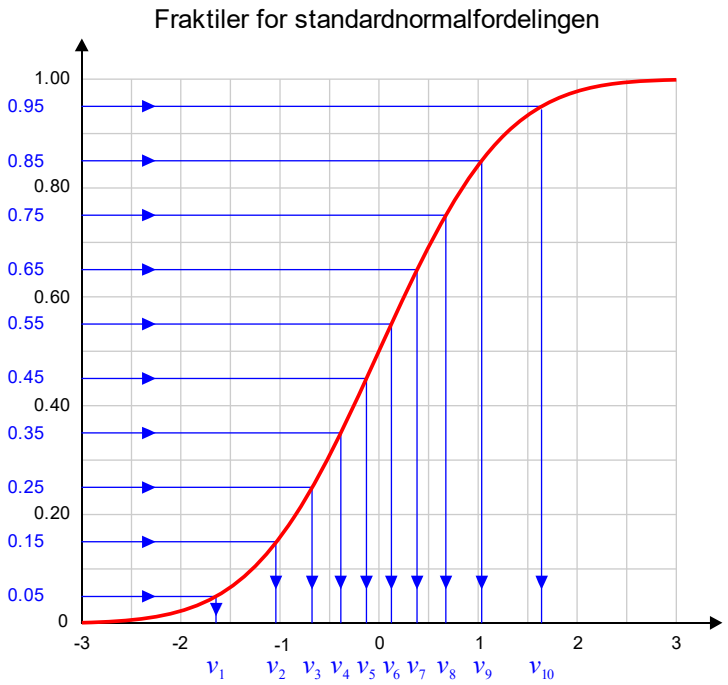
For ethvert p opfyldende $0 < p < 1$ definerer vi p-fraktilen for den empiriske fordelingsfunktion som den værdi x_p, som fremkommer ved at gå op til p på 2. aksen i plottet ovenfor, vandret hen til trappekurven og lodret ned på 1. aksen. Hvis den vandrette linje rammer den stiplede linje mellem to trin, så skal x_n være den værdi, som ligger umiddelbart under denne lodrette stiplede linje. Det er vist med et eksempel med blå linjer på figuren ovenfor. Hvis den vandrette linje derimod rammer direkte ind på et trin, er der tvetydighed. Så kan man vælge x-koordinaten for ethvert punkt på det vandrette trin, fx midtpunktet. Da vi imidlertid ikke får brug for det i det følgende, behøver vi ikke foretage

et valg her. For at afgøre, om stikprøven har en normalfordeling, vil vi afbilde et fraktilsæt for standardnormalfordelingen som funktion af det tilsvarende fraktilsæt for den empiriske fordelingsfunktion. Her skal man vælge sig en række p-værdier. Det er almindeligt at vælge lige så mange p-værdier som antallet af elementer i stikprøven, dvs. n, og man bruger ofte følgende:

(B2) $$p_i = \frac{i-\frac{1}{2}}{n},\ i = 1, 2, \ldots, n$$

I vores eksempel giver det følgende værdier: $p_1 = 0{,}05;\ p_2 = 0{,}15; \ldots; p_n = 0{,}95$.





På de to grafer på forrige side, er aflæst kvartilerne for $p_1, p_2, \ldots, p_{10}$ for henholdsvis den empiriske fordelingsfunktion og standardnormalfordelingen. Det giver de markerede fraktilsæt, henholdsvis $u_1, u_2, \ldots, u_{10}$ og $v_1, v_2, \ldots, v_{10}$. Bemærk i øvrigt, at sidstnævnte serie kan beregnes ved at anvende den inverse funktion til fordelingsfunktionen for standardnormalfordelingen på sandsynlighederne:

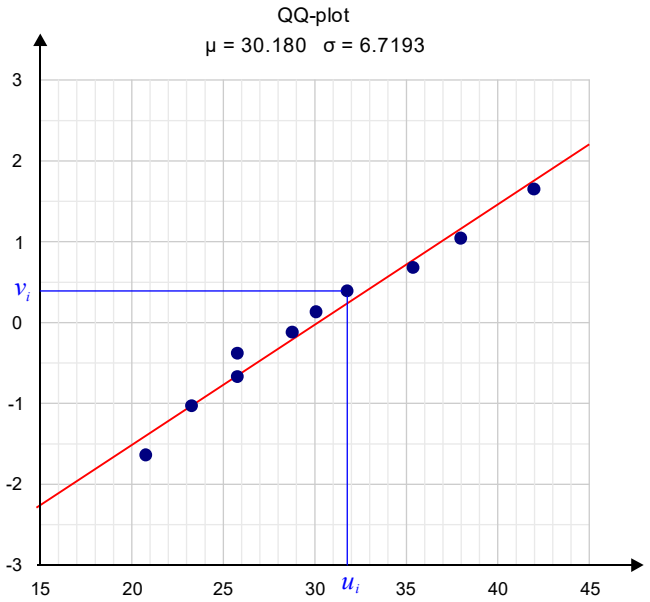
(B3) $$v_i = \Phi^{-1}(p_i)$$

QQ-plottet består da af punkterne $(u_1, v_1), (u_2, v_2), \ldots, (u_{10}, v_{10})$.

Lad os et øjeblik antage, at observationerne i stikprøven virkelig var omtrent normalfordelte med middelværdi μ og spredning σ, og at vi i stedet for standardnormalfordelingen havde anvendt den generelle normalfordeling med middelværdi μ og spredning σ i definitionen af QQ-plottet. Da ville punkterne i QQ-plottet ligge tæt på diagonalen $y = x$ i koordinatsystemet! Men heldigvis er der en snæver lineær sammenhæng mellem standardnormalfordelingen og den generelle normalfordeling. Vi har nemlig

(B4) $$X \sim N(\mu, \sigma) \Leftrightarrow Z \sim N(0,1) \quad \text{hvor } Z = \frac{X-\mu}{\sigma} = \frac{1}{\sigma} \cdot X - \frac{\mu}{\sigma}$$

Derfor kan vi nøjes med at anvende standardnormalfordelingen i definitionen af QQ-plottet. Med dette valg vil man normalt ikke få punkter omkring diagonalen, men om en anden linje. Det vil ofte være ret nemt visuelt at afgøre, om punkterne ligger omtrent på linje, og i det tilfælde er det en god indikation af, at man har at gøre med en normalfordeling. I øvrigt vil et QQ-plot også typisk indeholde linjen $y = \frac{1}{\sigma} \cdot x - \frac{\mu}{\sigma}$, hvor man har valgt den empiriske middelværdi $\bar{x}$ og stikprøvespredningen s som estimater for henh. μ og σ. Det er bedre end at udføre lineær regression på punkterne i QQ-plottet!



Bemærkning B2

Vores stikprøve var i det anvendte eksempel kun på 10 elementer – for at gøre situationen overskuelig. Det vil almindeligvis være for få observationer til at kunne afgøre om opsamlet data er normalfordelt. Tilfældigheder vil for nemt kunne spille ind.

Eksempel B3 (Normalfordelte residualer)

Hvis man i en konkret anvendelse ønsker at godtgøre, om man har at gøre med en *simpel lineær regressionsmodel*, så vil én af undersøgelserne typisk bestå i, at man undersøger, om residualerne er normalfordelte. Lad for eksempel $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ være punkterne i en stikprøve. Den tilhørende regressionslinje $y = a \cdot x + b$ bestemmes og herefter residualerne via formlen $r_i = y_i - (a \cdot x_1 + b)$. Man kan derefter lave et QQ-plot af residualerne $r_1, r_2, \ldots, r_n$ for at se, om de omtrent ligger på linje. Se eventuelt eksempel 5.16 i kapitel 5.

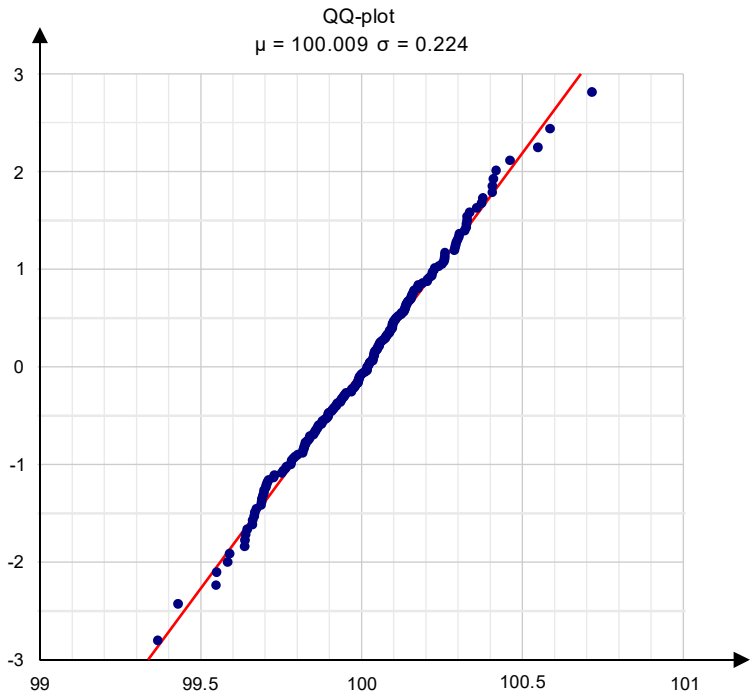
Eksempel B4 (Normalfordelte unøjagtigheder i produktionen)

I eksempel 3.11 i afsnit 3.3 blev det nævnt, at komponenter i produktionen ofte har unøjagtigheder, og at fejlene ofte er normalfordelte. I dette tilfælde er der tale om et firma, som producerer lodder til undervisningsbrug. Hensigten er at producere lodder med vægten 100 gram. I en kontrol bliver der udtrukket en stikprøve på 200 lodder med henblik på at undersøge variationen i loddernes vægt og om de er normalfordelte. En Excel fil med de 200 værdier for loddernes vægte sendes til en statistiker, som importerer data i sit CAS-værktøj.

Lodvægt (g)	99.824	100.006	...	100.094	99.695	99.927

Dele af data fra filen lodvaegte.xlsx

Heri laver personen et QQ-plot af lodvægtene, som giver resultatet herunder. Da punkterne ligger meget pænt på linje, godtager vi, at loddernes vægte er normalfordelte. Desuden får vi oplyst, at den empiriske middelværdi er 100,009 gram og at stikprøvespredningen er 0,224 gram.



Opgaver

1.1 Multiplikationsprincippet og additionsprincippet

Opgave 101

I en butik, der sælger softice, kan man vælge mellem vaffel eller bæger. Der kan vælges mellem vaniljeis eller farvet is. Som topping til hver af dem kan vælges enten guf, syltetøj eller krymmel. Lav et tælletræ, som viser mulighederne. Hvor mange er der?

Opgave 102

Løs følgende blandede opgaver:

a) Et firma producerer 4 typer mobiltelefoner, og hver af dem kan fås i 5 forskellige farver. Hvor mange forskellige mobiltelefoner kan man få hos producenten?

b) En bilproducent producerer 7 forskellige typer biler. Hver af dem fås i 3 forskellige farver, og hver kan fås med og uden klimaanlæg. Hvor mange forskellige biler kan man få?

Opgave 103

En 3g klasse skal vælge studenterhuer. Et firma tilbyder huen i en standardudgave og i en luksusudgave. Standardudgaven af huen fås med 3 forskellige skygger. Andet kan ikke varieres. Luksusudgaven derimod får med 5 forskellige skygger, med eller uden broderi samt med eller uden champagneglas. Benyt additionsprincippet og multiplikationsprincippet til at bestemme, hvor mange forskellige huer firmaet tilbyder?

Opgave 104

Der slås samtidigt med en grøn og en rød terning. På hvor mange forskellige måder kan man slå mindst én femmer? Terningerne regnes som forskellige, jf. farven.

Hjælp: Del evt. problemet op og brug additionsprincippet.
Antag først, at den grønne terning viser én femmer, dernæst den røde. Pas på overlap!

Opgave 105

Danske spil har et spil, der hedder *Tips 13*, hvor man skal gætte resultaterne i hver af i alt 13 fodboldkampe: Hjemmebanegevinst (1), uafgjort (x) eller udebanegevinst (2). Hvor mange muligheder er der total set? *Hjælp*: Benyt multiplikationsprincippet.

Opgave 106

Astrid, Bjørn, Cecilie, Dennis og Emilie skal fordeles i to forskellige rum.

a) På hvor mange måder kan de fem personer anbringes i rummene?

b) Samme spørgsmål, hvis der mindst skal være én person i hvert rum.

1.2 Permutationer og kombinationer

Opgave 107

På hvor mange måder kan man sætte 5 personer i rækkefølge?

Opgave 108

Til et svømmestævne skal 6 danske svømmere marchere ind i hallen sammen. På hvor mange måder kan de sættes i rækkefølge?

Opgaver 109

Udregn nedenstående kombinationer manuelt ved hjælp af formlen (3) i sætning 1.11. (Foretag en reduktion af brøkerne, før du ganger ud):

a) $K(4,3)$ b) $K(6,4)$ c) $K(5,2)$ d) $K(20,2)$ e) $K(6,0)$

Opgave 110

Udregn nedenstående permutationer manuelt ved hjælp af formlen (1) i sætning 1.10. (Foretag en reduktion af brøkerne, før du ganger ud):

a) $P(5,2)$ b) $P(7,1)$ c) $P(6,3)$

Opgave 111

Udregn nedenstående kombinationer og permutationer ved hjælp af dit CAS-værktøj.

a) $K(13,7)$ b) $K(75,65)$ c) $K(25,5)$ d) $P(12,5)$ e) $P(10,7)$

Opgave 112

Tre bogstaver skal udtrækkes fra følgende mængde af bogstaver: a, b, c, d og e.

a) Hvor mange måder kan man udtrække de tre bogstaver på, hvis rækkefølgen er ligegyldig? Opskriv alle kombinationer.
b) Hvor mange måder kan man udtrække de tre bogstaver på, hvis rækkefølgen *ikke* er ligegyldig.

Opgave 113

Bevis at der gælder følgende generelle relation: $K(n,r) = K(n,n-r)$, idet du bruger formel (3) fra sætning 1.11. Forsøg desuden at redegøre for, hvorfor det er ret klart intuitivt, når man tænker på at trække noget ud …

Opgave 114

Fire elever fra en klasse med 17 elever skal udtrækkes med henblik på rengøring efter fest næste dag. På hvor mange måder kan de fire elever udvælges?

Opgave 115

I en forening er der møde i bestyrelsen, som består af 16 personer. Der skal udnævnes en formand, en næstformand, en kasserer, foruden en person, som skal tage sig af kontakten til medierne. Ingen personer melder sig selv til posterne. På hvor mange måder kan man besætte de fire poster?

Opgave 116

Der er en kasse med i alt 15 kugler i, hvoraf de 6 er blå og de 9 røde. På hvor man måder kan man sætte de 15 kugler i rækkefølge? En kombination er vist på figuren herunder.
Hjælp: "Udtræk" de positioner, hvor der skal være en blå kugle …

Opgave 117 (Poker)

Vi ønsker at regne på flere poker-hænder i stil med dem i eksempel 1.14. Der uddeles 5 kort én gang fra et spil med 52 kort. De fem kort omtales som "en hånd".

a) Vis at antallet af mulige hænder med 3 ens er 54912.
b) Vis at antallet af mulige hænder med 2 ens er 1098240.

Hjælp: Tænk på fremgangsmåden i eksempel 1.14 b).

Opgave 118 (Poker)

Hvor mange måder kan man få en hånd i poker (5 kort) bestående af tre 8'ere og to esser?
Hjælp: Bemærk her, at i forhold til eksempel 1.14 *er* talværdierne i denne hånd med fuldt hus allerede fastlagt, så kun kulørerne skal vælges!

Opgave 119 (Kortspil)

Der uddeles 4 kort fra et almindeligt kortspil med 52 kort. På hvor mange måder kan man få to 9'ere og to billedkort?

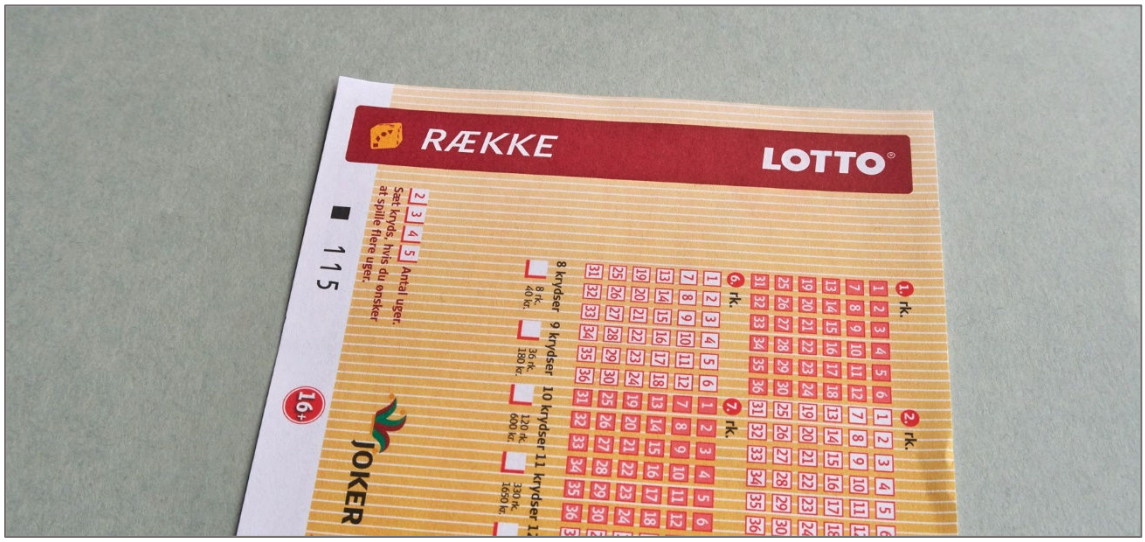
Opgave 120 (Stikprøver)

I en population på 70 aber skal udtrækkes 8 aber til en test. Hvor mange måder kan stikprøven udtrækkes på?

Opgave 121 (Stikprøver)

I en virksomhed er der ansat 18 mænd og 11 kvinder. På hvor mange måder kan man på tilfældig vis udtrække en gruppe på 4 mænd og 3 kvinder til en bestyrelse? *Hjælp*: Husk at det her er "både og".

Opgave 122 (Lotto)



I Lotto under *Danske spil* udtrækkes 7 tal blandt tallene fra 1 til og med 36.

a) Hvor mange mulige udtrækninger er der?

Det er klart, at for at få alle 7 rigtige, skal man ramme præcist den kombination, som blev udtrukket. For at få 6 rigtige, behøver man ikke have alt rigtigt. Derfor er der flere kombinationer af tal, som giver 6 rigtige.

b) Vis, at der er 203 forskellige valg af 7 forskellige tal, som vil give 6 rigtige. *Hjælp*: Der er 7 rigtige og 29 forkerte. Du skal vælge 6 rigtige og én forkert.

2.2 Endeligt sandsynlighedsfelt

Opgave 201

Ved slag med en falsk terning har man nedenstående sandsynlighedsfordeling for terningens visning, hvor sandsynligheden for udfaldet 6 dog er ukendt.

u	1	2	3	4	5	6
$P(u)$	0,21	0,10	0,15	0,22	0,13	p

Bestem den manglende sandsynlighed p.

Opgave 202

Når man trykker på en knap i en app på mobiltelefonen får man en af udfaldene: *a*, *b*, *c*, *d* og *e*, med de sandsynligheder, som fremgår af tabellen nedenfor. Sandsynligheden for udfaldet *c* er dog ukendt.

u	a	b	c	d	e
$P(u)$	0,3	0,1	p	0,1	0,2

a) Bestem den manglende sandsynlighed p.
b) Bestem sandsynlighederne for hændelserne $H=\{a,b,c\}$ og $G=\{a,d,e\}$.
c) Bestem sandsynligheden for hændelserne $H\cup G$ og $H\cap G$.
d) Tegn et stolpe- eller søjlediagram for sandsynlighedsfordelingen.

Opgave 203 (Endeligt sandsynlighedsfelt)

I en kasse er der 3 røde kugler, 7 grønne kugler, 4 blå kugler og 6 sorte kugler. I et eksperiment trækkes der på tilfældig måde en enkelt kugle fra kassen.

a) Bestem eksperimentets udfaldsrum U.
b) Angiv eksperimentets sandsynlighedsfunktion P, dvs. udregn sandsynlighederne for de enkelte udfald og skriv dem ind i en tabel (sandsynlighedsfordelingen).
c) Giv et eksempel på en mulig hændelse H, og bestem dens sandsynlighed.

Opgave 204 (Træk fra kortspil)

Fra et kortspil med 52 kort trækkes på tilfældig vis ét kort.
a) Bestem sandsynligheden for, at kortet er en ruder.
b) Bestem sandsynligheden for, at kortet er en billedkort.
c) Bestem sandsynligheden for, at kortet er *både* et billedkort *og* en ruder.
d) Bestem sandsynligheden for, at kortet er *enten* et billedkort *eller* en ruder.
e) Bestem sandsynligheden for, at kortet *ikke* er et billedkort.

Opgave 205 (Kast med én terning)

Der kastes med en enkelt terning. Bestem sandsynligheden for følgende hændelser:
a) Terningen viser et *ulige* antal øjne.
b) Terningen viser mindst 3 øjne.
c) Terningen viser hverken 5 eller et lige antal øjne.

Opgave 206 (Symmetrisk sandsynlighedsfelt)

I et eksperiment er der otte mulige udfald: 1, 2, 3, 4, 5, 6, 7 og 8. Det oplyses, at der er tale om et symmetrisk sandsynlighedsfelt. Bestem sandsynligheden for følgende hændelse: $H=\{2,4,5\}$.

Opgave 207 (Møntkast)

Der kastes en mønt tre gange, og man interesserer sig for, om der i hvert kast kommer en *krone* eller en *plat*. Find en smart måde at angive et udfald ved de tre kast med en mønt (se evt. eksempel 2.9). Bestem udfaldsrummet og opskriv sandsynlighedsfordelingen.

Opgave 208 (Møntkast)

I fortsættelse af opgave 207, hvor der kastes 3 gange med en mønt, og hvor man interesserer sig for kombinationen af *krone*/*plat* i hvert kast: Hvad er sandsynligheden for følgende hændelser:

a) At få plat i første kast?
b) At få plat i både første og sidste kast?
c) At få mindst en krone?

Hjælp: Opskriv først de udfald, som der er i hver hændelse, og udregn derefter sandsynlighederne.

Opgave 209 (Kast med to terninger)

Vi betragter kast med to terninger – en grøn og en rød – ligesom i eksempel 2.8.

a) Indtegn følgende hændelser i kvadratet til højre:
   H : Summen af øjnene er 10.
   G : Den grønne terning viser mindst 5.
b) Bestem sandsynlighederne for de to hændelser.
c) Bestem sandsynligheden for hændelserne:
   $H \cup G$ og $H \cap G$.

Rød terning	Grøn terning: 1	2	3	4	5	6
6	(1,6)	(2,6)	(3,6)	(4,6)	(5,6)	(6,6)
5	(1,5)	(2,5)	(3,5)	(4,5)	(5,5)	(6,5)
4	(1,4)	(2,4)	(3,4)	(4,4)	(5,4)	(6,4)
3	(1,3)	(2,3)	(3,3)	(4,3)	(5,3)	(6,3)
2	(1,2)	(2,2)	(3,2)	(4,2)	(5,2)	(6,2)
1	(1,1)	(2,1)	(3,1)	(4,1)	(5,1)	(6,1)

Opgave 210 (Udtrækning)

I en skål er der 30 røde bolsjer og 15 gule bolsjer. Oscar trækker på tilfældig vis og uden at kigge 3 bolsjer fra skålen.

a) Hvad er sandsynligheden for, at han får to røde og et gult bolsje?
b) Hvad er sandsynligheden for, at han får tre gule bolsjer?

Hjælp: Udregn antal gunstige og antal mulige udfald, idet du bruger $K(n,r)$. Gunstige i a): der skal trækkes to røde bolsjer ud af 30 og et gult bolsje ud af 15 … (både og).

Opgave 211 (Poker)

Bestem sandsynligheden for at få 3 ens i Poker.

Hjælp: Ligesom i eksempel 2.13 kan du bestemme antal gunstige udfald (se evt. opgave 117) og dividere med antal mulige udfald.

Opgave 212 (Udtrækning)

Isabella og Ulla er veninder og går i samme klasse med i alt 23 elever. Blandt eleverne udtrækkes på tilfældig vis 4 elever, som skal stå for madlavningen til en fest.

a) På hvor mange måder kan man udtrække de fire elever?
b) Hvad er sandsynligheden for, at både Isabella og Ulla kommer på madholdet?

Opgave 213 (Udtrækning)

Det er svært ved et øjekast at afgøre, hvilket køn en kanin har. En population indeholder 50 kaniner, hvor det vides, at de 30 er hun-kaniner og de 20 er han-kaniner. Der udtrækkes på tilfældig måde 10 kaniner i en stikprøve.

a) På hvor mange måder kan man udtrække 10 kaniner af populationen?
b) Bestem sandsynligheden for, at stikprøven indeholder 6 han- og 4 hun-kaniner.

Opgave 214 (Udtrækning)

I en lille Tombola kan man trække kugler fra en kasse. Kuglerne kan indeholde en gevinst. 5 af kuglerne indeholder en gevinst på 100 kr., 10 kugler indeholder en gevinst på 50 kr., mens de resterende 20 kugler er nitter (ingen gevinst). Michael køber i alt 6 kugler. Der er ingen tilbagelægning!

a) På hvor mange måder kan man udtrække de 6 kugler?
b) Bestem sandsynligheden for, at Michael får 1 kugle med 100 kr. gevinst, 3 kugler med 50 kr. gevinst og 2 nitter.

Opgave 215 (Udtrækning)

Man ved af erfaring, at 10% af boltene til noget teknisk udstyr er defekte. En kasse indeholder 40 bolte, hvoraf 4 er defekte.

a) Der udtrækkes på tilfældig vis 10 bolte fra kassen. Hvad er sandsynligheden for, at mindst én af dem er defekte? *Hjælp*: Udregn den komplementære hændelse, som er at, *ingen* er defekte, og brug derefter sætning 2.7 c).
b) Hvor mange bolte skal personen udtrække, før sandsynligheden for mindst én defekt bolt overstiger 90%. *Hjælp*: Du kan prøve dig frem eller måske opstille en ligning …

Opgave 216 (Udtrækning)

I en børnehave er der 12 piger og 8 drenge. Der skal ved tilfældig lodtrækning udvælges fire børn, som skal stå for at pynte op til en fødselsdag.

a) På hvor mange måder kan de fire børn udvælges?
b) Bestem sandsynligheden for, at der er mindst 2 piger med i gruppen.

Hjælp: Mindst 2 piger betyder, at der må være *enten* 2, 3 eller 4 piger. Du skal udregne sandsynligheder hver for sig og lægge sammen (se sætning 2.7 d)).

Opgave 217 (Flere kast med terning - uafhængighed)

I det følgende skal du udnytte *uafhængighed* ved kast med terning.

a) Der kastes med én terning to gange efter hinanden. Bestem sandsynligheden for få netop to femmere.
b) Hvad er sandsynligheden for ved et enkelt kast med en terning at få en ikke-femmer? Udnyt sætning 2.7 c).
c) Hvad er sandsynligheden for ved 5 kast med en terning at få først 2 femmere og derefter 3 ikke-femmere? Undertiden skriver vi hændelsen således: $(5, 5, \overline{5}, \overline{5}, \overline{5})$.

NB! Når vi i næste kapitel kommer til den såkaldte *binomialfordeling*, får vi brug for at beregne sandsynligheder af typen i c).

Opgave 218 (Uafhængighed)

I eksempel 2.17 så vi, at hændelserne A og B er uafhængige. Nu ændrer vi en smule på hændelsen B, mens A er uændret:

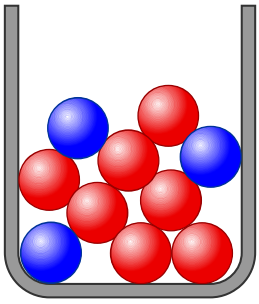
A : "Den grønne terning viser 2". B : "Summen af terningerne viser 6".

Vis, at disse to hændelser *ikke* er uafhængige.

Hjælp: Vis at (3) i definition 2.15 ikke er opfyldt. Indtegn hændelserne i det sædvanlige kvadrat (jf. eksempel 2.8) for at bestemme de tre sandsynligheder.

Opgave 219 (Udtrækning med og uden tilbagelægning)

På tilfældig vis og med ryggen til trækkes der tre gange efter hinanden en kugle fra en krukke, som fra start indeholder 3 blå kugler og 7 røde kugler.



a) Hvad er sandsynligheden for at få en rød kugle i alle tre trækninger, når kuglen lægges tilbage efter hver trækning?
b) Hvad er sandsynligheden for at få en rød kugle i alle tre trækninger, hvis kuglen *ikke* lægges tilbage efter hver trækning?

Hjælp: a) Udnyt eventuelt uafhængighed. b) Udnyt det symmetriske sandsynlighedsfelt og bestem antal gunstige og antal mulige udfald ved de tre udtrækninger.

Opgave 220 (Fødselsdagsproblemet)

a) Hvad er sandsynligheden for, at mindst to elever i en klasse med 20 elever har fødselsdag på samme dag? (se eksempel 2.14).
b) Hvor mange elever skal der være i klassen, før sandsynligheden overstiger 50%.

Opgave 221 (Poker)

Bestem sandsynligheden for at få 4 ens i poker.

Opgave 222 (Le Chevalier de Mérés problem)

Den franske adelsmand og gambler *Le Chevalier de Méré* alias *Antoine Gombaud* (1607-1684) studerede følgende to hændelser:

1) At få mindst én sekser ved fire kast med *en* terning.
2) Mindst én gang at få en dobbelt-sekser ved 24 kast med *to* terninger.

Selv om han havde en anelse om, hvilken hændelse, der var den mest sandsynlige, kunne de Méré ikke redegøre for det. Derfor henvendte han sig til den store franske matematiker *Blaise Pascal* (1623 – 1662). Pascals svar bekræftede Mérés formodning. I det følgende skal du prøve selv at bestemme de to sandsynligheder. *Hjælp*: Se på den komplementære hændelse og benyt sætning 2.7 c). Benyt desuden uafhængighed.

Opgave 223 (Hændelser og Venn-diagrammer)

Betragt mængdeoperationerne samt Venn-diagrammerne side 18. Du skal vise de to identiteter nedenfor, hvor A og B er vilkårlige delmængder af udfaldsrummet U. *Hjælp*: Vis, at hvis et element u tilhører venstresiden, så vil det også tilhøre højresiden, og omvendt. Alternativt: skraver i Venn-diagrammer, startende med to overlappende mængder A og B. Skraver mængderne på højre og venstre side. Giver det det samme?

a) Vis, at $A^c \cup B^c = (A \cap B)^c$. b) Vis, at $A^c \cap B^c = (A \cup B)^c$.

2.3 Stokastisk variabel

Opgave 224

Betragt det eksperiment, hvor der kastes med en enkelt terning. Betragt den stokastiske variabel X: Det øjnene viser. Bestem følgende sandsynligheder:

a) $P(X = 3)$ b) $P(X \le 2)$ c) $P(X \ge 4)$

idet du først gør dig klart hvilke udfald, som hver af følgende hændelser indeholder: $X = 3$ og $X \le 2$ og $X \ge 4$.

Opgave 225

En stokastisk variabel X har nedenstående sandsynlighedsfordeling, hvor p er ukendt.

x	–3	–2	1	2	3	4
$P(X = x)$	0,21	0,16	p	0,32	0,05	0,17

a) Bestem først den ukendte parameter p og derefter følgende sandsynligheder: $P(X > 0)$, $P(2 \le X \le 4)$ og $P(X \le 2)$.

Opgave 226

En stokastisk variabel X har nedenstående sandsynlighedsfordeling.

x	0	5	10	15	20
$P(X=x)$	0,55	0,10	0,20	0,10	0,05

Bestem de ukendte sandsynligheder: $P(X \le 10)$, $P(X \le 5)$ og $P(X \ge 5)$.

Opgave 227

En stokastisk variabel kan antage de fire værdier 1, 2, 3 og 4. Følgende oplysninger gives: $P(X \le 2) = 0{,}45$; $P(X \ge 2) = 0{,}75$; $P(X \le 3) = 0{,}85$. Bestem herudfra sandsynligheden for hver af de fire værdier af X, altså sandsynlighedsfordelingen for X.

Opgave 228

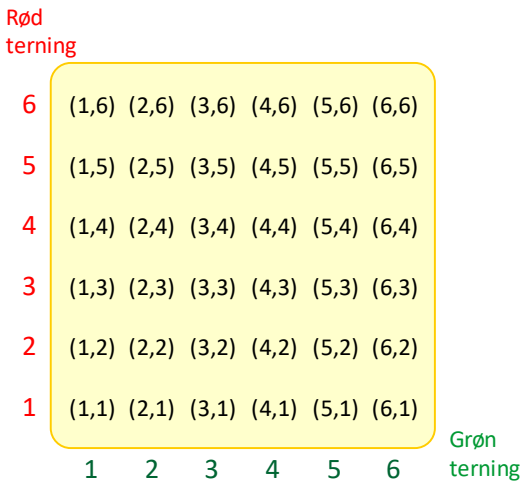
Eksperiment: Der trækkes to tal fra mængden $\{1, 2, 3, 4\}$.
Den stokastiske variabel X: angiver den største værdi af de to udtrukne tal. Bestem de mulige værdier X kan antage, og bestem derefter sandsynlighedsfordelingen for X.

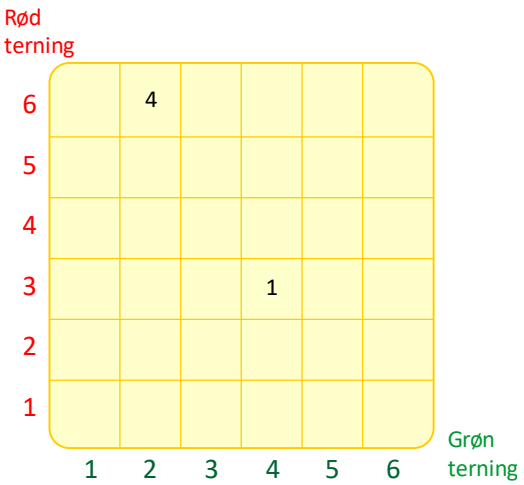
Opgave 229 (Kast med to terninger)

Betragt følgende eksperiment: Et kast med to terninger, en grøn og en rød. Antal øjne betragtes. Lad den stokastiske variabel X angive *forskellen* på øjnene af de to terninger.

a) Hvilke værdier kan X antage? Bestem sandsynlighedsfordelingen for X.
b) Bestem sandsynligheden for, at forskellen på terningernes øjne er 2, dvs. $P(X = 2)$.
c) Bestem sandsynligheden $P(X \le 2)$. Udtryk denne sandsynlighed sprogligt med ord, lige som under spørgsmål b).

Hjælp: Benyt eventuelt kvadraterne nedenfor og indtegn på det højre kvadrat værdien af X for hvert udfald i kvadratet – i stil med eksempel 2.22. To eksempler er givet.





Opgave 230 (Møntkast)

Eksperimentet er et kast med tre mønter. Man interesserer sig for, om en mønt viser plat eller krone. Det er en pædagogisk hjælp at tænke på, at mønterne er nummererede. Lad for eksempel (k, p, k) betyde, at mønt 1 viser krone, mønt 2 viser plat og mønt 3 krone.

a) Opskriv de 8 mulige udfald. Overvej hvorfor de er lige mulige?

I det følgende lader vi X være den stokastiske variabel, som angivet antallet af plat.

b) Angiv de mulige værdier for X og bestem sandsynlighedsfordelingen for X.

2.4 Middelværdi, varians og spredning

Opgave 231

En stokastisk variabel X kan antage værdierne 0, 1, 2, 3, 4 og 5 og har følgende sandsynlighedsfordeling:

x	0	1	2	3	4	5
$P(X=x)$	0,15	0,10	0,30	0,10	0,20	0,15

a) Bestem middelværdien $E(X)$.

b) Bestem variansen $Var(X)$ og spredningen $\sigma(X)$.

Opgave 232

En stokastisk variabel X har sandsynlighedsfordelingen vist herunder, hvor p er en ukendt sandsynlighed.

x	–2	4	5	8	10	15
$P(X=x)$	0,07	0,18	0,32	0,14	p	0,12

a) Bestem den ukendte sandsynlighed p.

b) Bestem middelværdien $E(X)$ samt variansen $Var(X)$.

Opgave 233

En stokastisk variabel X har sandsynlighedsfordelingen vist herunder, hvor k er en ukendt parameter. Det oplyses, at middelværdien for X er 1,76.

x	–5	2	k	8
$P(X=x)$	0,20	0,32	0,43	0,05

a) Bestem den ukendte parameter k og tegn et stolpediagram over fordelingen.

Opgave 234 (Spil i Casino)

I et spil i et Casino foretager spilleren en indsats på 100 kr. De mulige gevinster i kroner er: 0, 150, 200 og 500. Efter indsatsen er trukket fra, er gevinsterne set fra spillerens synspunkt (regnet med fortegn) altså henholdsvis -100, 50, 100 og 400 kroner. Vi lader den stokastiske variabel X angive den reelle gevinst ved et spil. Sandsynlighederne for de enkelte reelle gevinster fremgår af tabellen nedenfor.

x	-100	50	100	400
$P(X=x)$	0,50	0,35	0,12	0,03

a) Hvor meget vil spilleren tabe i gennemsnit pr. spil? *Hjælp*: Udregn $E(X)$.

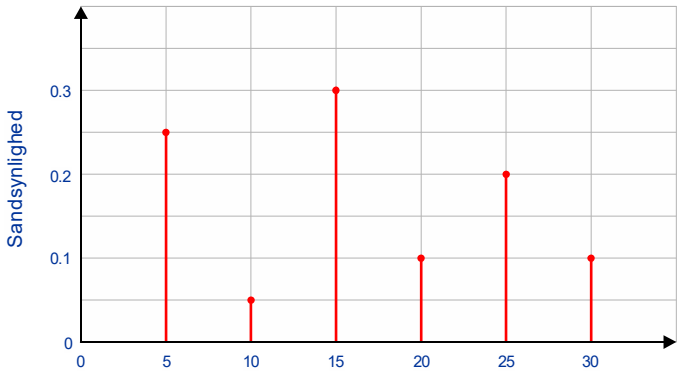
b) Bestem variansen $Var(X)$.

Bankøren ønsker at øge sin profit til 15%, altså så han i snit vinder 15 kr. pr. spil. For at opnå det, overvejer bankøren kun at justere på de to første sandsynligheder i tabellen.

c) Hvad skal bankøren sætte de to første sandsynligheder til for at få den ønskede profit i gennemsnit pr. spil?

Opgave 235

Sandsynlighedsfordelingen for en stokastisk variabel X er afbildet med et stolpediagram nedenfor. Bestem middelværdi, varians og spredning for X.



Opgave 236 (Gevinst ved spil)

En bookmaker foreslår følgende spil til en spiller: Ét spil består i at kaste tre gange med en mønt. Hvis spilleren udelukkende får plat, skal denne betale 20 kr. Hvis spilleren får krone to gange lige efter hinanden, vinder denne 5 kr. I alle andre tilfælde vinder spilleren 1 kr. Er det et fornuftigt spil for spilleren i det lange løb?

Hjælp: Opskriv først de 8 mulige udfald ved ét spil, dvs. tre kast – husk at rækkefølgen er væsentlig! Indfør en stokastisk variabel X, som skal være gevinsten ved ét spil. Hvilke værdier kan X antage? Bestem sandsynligheden for hver af disse værdier ved at betragte listen med de 8 mulige udfald. Da du således har sandsynlighedsfordelingen for X, kan du afgøre spørgsmålet ved at udregne middelværdien $E(X)$.

Opgave 237 (Skrabelod)

På MINI QUICK skrabelodderne vinder man beløbet, hvis der er tre ens. Antallet af præmier er specificeret på bagsiden af loddet:

5	×	100000 kr.
3000	×	1000 kr.
10000	×	500 kr.
30000	×	100 kr.
50000	×	50 kr.
113560	×	25 kr.
786100	×	10 kr.

Det oplyses, at der i alt er 4950000 lodder.

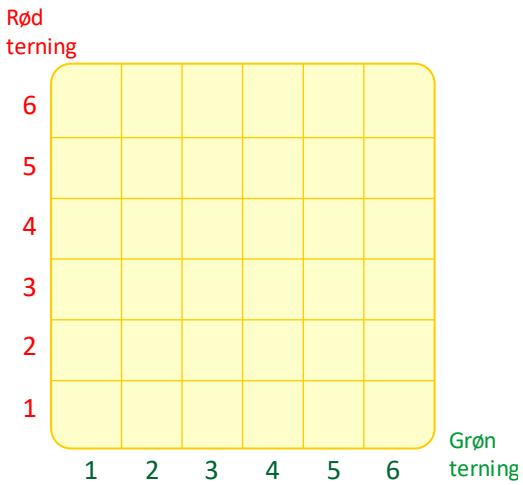
a) Hvor mange lodder er nitter, dvs. hvor mange giver ingen gevinst?
b) Hvor stor en brøkdel af lodderne er der gevinst på?

Indfør nu en stokastisk variabel X, som angiver gevinsten ved køb af et skrabelod. Her skal man huske, at selv om man vinder på sit lod, så har man stadig betalt 10 kr. for loddet. Derfor vælger vi at trække 10 kr. fra beløbene, så de mulige gevinster (fraregnet betalingen) er følgende, regnet i kroner: –10, 0, 15, 40, 90, 490, 990, 99990.

c) Udregn sandsynlighedsfordelingen for X.
d) Udregn middelværdien af gevinsten ved hvert lod, dvs. udregn $E(X)$.

Opgave 238 (Gevinst ved spil med to terninger)

I stil med eksempel 2.22 og eksempel 2.25, kigger vi igen på et spil, hvor der kastes med en grøn og en rød terning. Den stokastiske variabel X angiver gevinsten ved et spil, set fra spillerens synspunkt. Spilreglerne er følgende: Hvis spilleren slår mindst én sekser, så skal denne betale 10 kr. Hvis begge terninger viser et ulige antal øjne, så modtager spilleren produktet af det antal øjne, terningerne viser. Hvis spilleren slår (3,5), så vinder spilleren for eksempel $3 \cdot 5 = 15$ kroner. Ved alle andre slag er der hverken tab eller gevinst.



a) Bestem sandsynlighedsfordelingen for X. Benyt eventuelt det tomme kvadrat til højre til at plotte gevinster ind (regnes positiv hvis spilleren vinder, og negativt, hvis denne taber).
b) Beregn middelværdien $E(X)$. Hvor meget vil spilleren i gennemsnit miste i hvert spil?

Opgave 239 (Funktion af stokastiske variable – teoretisk projekt)

Vi skal først se på lidt teori i denne opgave. Man kan også betragte stokastiske variable, som er funktioner af andre stokastiske variable: $Y = a \cdot X + b$ er således et eksempel på en stokastisk variabel Y, som er en lineær funktion af den stokastiske variabel X. For at få sandsynlighedsfordelingen for Y, finder man først ud af, hvilke værdier Y kan antage. I eksemplet nedenfor er $Y = 2X + 3$, så vi indsætter alle de mulige værdier for X i funktionen $f(x) = 2x + 3$, hvilket giver $f(-5) = -7$, $f(3) = 9$, $f(5) = 13$ og $f(8) = 19$. Sandsynlighederne for hver af disse y-værdier er de samme, som for de tilhørende x-værdier. Dette skyldes, at der ikke er gengangere i y-værdierne. Havde der været gengangere, måtte vi have slået nogle af sandsynlighederne sammen (Overvej!).

x	–5	3	5	8
$P(X = x)$	0,15	0,40	0,20	0,25

y	–7	9	13	19
$P(Y = y)$	0,15	0,40	0,20	0,25

a) Bestem middelværdien $E(X)$ for X og middelværdien $E(Y)$ for Y. Vis desuden, at der gælder $E(Y) = 2 \cdot E(X) + 3$.

Sidstnævnte egenskab i spørgsmål a) er ikke en tilfældighed. Der gælder nemlig følgende:

Sætning 1

Lad X være en stokastisk variabel og lad $Y = a \cdot X + b$, hvor a og b er konstanter. Da er Y en stokastisk variabel med $E(Y) = a \cdot E(X) + b$ og $Var(Y) = a^2 \cdot Var(X)$.

b) Prøv at bevise første påstand i sætning 1 angående middelværdien $E(Y)$, idet du benytter definition 2.24 side 30. *Hjælp*: Husk at værdierne for Y er $y_i = a \cdot x_i + b$.

Som bekendt kan man omregne temperaturer i grader Celcius til temperaturer i grader Fahrenheit via formlen $y = 1{,}8 \cdot x + 32$. Antag, at man havde en fordeling af temperaturer i °C og havde udregnet middelværdien til 17,5°C, men egentligt ønskede middelværdien af temperaturerne i °F. Da ville det ifølge sætning 1 ikke være nødvendigt først at omregne alle de oprindelige data til °F og derefter udregne middelværdien af disse. Man kan lige så godt proppe middelværdien på 17,5°C direkte ind i formlen og få middeltemperaturen i °F: $1{,}8 \cdot 17{,}5 + 32 = 63{,}5$. Måske ikke så overraskende, men lineariteten af middelværdien har også stor betydning i teoretiske overvejelser. Der gælder en anden smuk egenskab, som vi dog ikke skal bevise:

Sætning 2

Lad X være en stokastisk variabel. Da gælder følgende om variansen:

$$Var(X) = E(X^2) - (E(X))^2 = E(X^2) - \mu^2$$

c) Vis at formlen i sætning 2 holder stik for X i begyndelsen af denne opgave.

Hjælp: Bemærk, at du først må finde sandsynlighedsfordelingen for den stokastiske variabel $Y = X^2$. NB! Her forekommer overlap, idet to forskellige x-værdier giver samme y-værdi. Derfor er der kun tre mulige værdier af Y.

Formlen i sætning 2 er ofte en smule lettere at bruge end den oprindelige definition 2.24 side 30. Igen har sætning også en teoretisk betydning.

Opgave 240 (Filmvurdering)

Man har foretaget en stor undersøgelse af, hvor godt danskerne kan lide en ny dansk film. De udspurgte er blevet bedt om at give filmen points på en skala fra 1 til 5. Undersøgelsen resulterede i denne frekvensfordeling:

x	1	2	3	4	5
$P(X = x)$	0,05	0,12	0,43	0,31	0,09

Lad X betegne en stokastisk variabel, som angiver hvor mange points en tilfældig udspurgt person, der har set filmen, vil give den.

a) Bestem middelværdien af X. Hvad fortæller den sagt med ord?
b) Hvad er sandsynligheden for, at en tilfældigt udspurgt person, som har set filmen, vil tildele den mindst 4 points? Samme spørgsmål med højst 2 points.
c) Tegn et stolpe- eller pindediagram over fordelingen.

Opgave 241

Nogle børn i en børnehave har lavet et lykkehjul med de mulige udfald a, b, c, d, og e. Gevinsterne for de enkelte udfald samt deres frekvenser er afbildet i tabellen her, hvor der dog er en ukendt sandsynlighedsparameter p.

Udfald	a	b	c	d	e
Gevinst	–10	1	5	10	25
Sandsynlighed	0,40	p	0,20	0,10	0,05

a) Bestem den ukendte sandsynlighed p.
b) Bestem middelværdien $E(X)$.

Børnene finder du af, at de ikke helt opnår det udbytte, som de ønsker. Derfor vil de sætte den største gevinst ned, så de i gennemsnit tjener 1 kr. pr. spil

c) Hvad skal de sætte den største gevinst for udfald e ned til? (hele antal kr.).

Opgave 242 (Odds i spil – projekt)

Som bekendt sætter bookmakere ofte odds på deres spil. I denne opgave skal vi se på selve begrebet *odds*. Der findes flere forskellige definitioner på odds, som det fremgår af skemaet 2 sider fremme. I det følgende vil vi lade det være underforstået, at det er den *europæiske* definition på odds, vi har med at gøre. Hvis et betting firma sætter odds 4,0 betyder det, at man får indsatsen igen 4 gange, hvis man rammer plet, ellers taber man indsatsen. Men det er klart, at firmaerne skal tjene på deres spil, så de sætter odds lavere end de ville, hvis de ikke skulle tjene på spillet.

Den teoretiske europæiske definition på odds er $odds = 1/p$, altså den reciprokke værdi af sandsynligheden p. Her er det underforstået, at p er sandsynligheden for, at det resultat, bookmakeren sætter odds på, indtræffer. Sådanne sandsynligheder bliver beregnet af eksperter, som kigger på statistikker og generelt set har en god fingerspidsfornemmelse for, hvad der rører sig. Lad os indføre en stokastisk variabel med henblik på at afgøre, om det er en økonomisk gevinst for bookmakeren at udbyde et bestemt spil.

Eksperiment: Der gennemføres en kamp eller et spil, hvor en spiller har sat 100 kr. på et bestemt resultat. De to mulige udfald er, om resultatet indtræffer eller det ikke indtræffer. Sandsynligheden for, at resultatet indtræffer, antages at være $p = 25\% = 0{,}25$. Det antages desuden, at bankøren benytter odds givet ved formlen $odds = 1/p = 1/0{,}25 = 4$.

Den stokastiske variabel X: Angiver gevinsten ved kampen/spillet set fra bookmakerens synspunkt, dvs. hvis bookmakeren vinder regnes positivt, ellers negativt.

a) Argumenter for, at sandsynlighedsfordelingen for X ser ud som nedenfor.

x	–300	100
$P(X = x)$	0,25	0,75

b) Vis, at middelværdien af den stokastiske variabel X er 0, dvs. at bookmakeren i det lange løb hverken vinder eller taber ved at udbyde spillet.

Det er klart, at bookmakeren ønsker at have profit af sin virksomhed, så han må sætte odds ned. Det kan for eksempel ske indirekte ved at indføre en fiktiv sandsynlighed og så stadig bruge formlen for odds. Han vælger at *fremskrive* sandsynligheden for, at spilleren vinder, med $k = 1{,}08$, så den bliver $\tilde{p} = k \cdot p = 1{,}08 \cdot 0{,}25 = 0{,}27$. Ifølge formlen for odds fås dermed de opdaterede odds: $\widetilde{odds} = 1/\tilde{p} = 1/0{,}27 = 3{,}7037$. De korrekte sandsynligheder for gevinst og tab ændrer sig naturligvis ikke, men det gør odds og dermed gevinst.

c) Argumenter for, at sandsynligheden for X nu kommer til at se således ud:

x	–270,37	100
$P(X = x)$	0,25	0,75

d) Bestem middelværdien for X. Hvor meget vil bookmakeren nu vinde i gennemsnit pr. spil? Samme spørgsmål i procent af indskuddet?

I praksis er der sikkert ikke en helt håndfast procedure hos bookmakerne, fordi de også skal se attraktive ud i forhold til de andre betting firmaer. Derfor er der formentlig også nogle ad hoc ting inde over i praksis. Vi vil nu tage skridtet og generalisere ovenstående til formler. Spillerens indskud betegnes med b, sandsynligheden for at resultatet indtræffer betegnes p og odds beregnes via formlen $odds = 1/p$. Den stokastiske variabel X angiver gevinsten ved kampen/spillet set fra bookmakerens synspunkt, regnet med fortegn.

e) Argumenter for, at sandsynlighedsfordelingen for X ser således ud, idet du gør det samme som ovenfor, nu blot med variable:

x	$-\left(\frac{1}{p}-1\right)\cdot b$	b
$P(X = x)$	p	$1-p$

f) Vis at middelværdien for X er lig med 0, ligesom i b).

For at kunne opnå en profit, fremskriver vi igen sandsynligheden p med en faktor k, så vi får en slags fiktiv sandsynlighed $\tilde{p} = k \cdot p$ samt opdaterede odds: $\widetilde{odds} = 1/\tilde{p}$.

g) Argumenter for, at sandsynligheden for X nu kommer til at se således ud:

x	$-\left(\frac{1}{k\cdot p}-1\right)\cdot b$	b
$P(X = x)$	p	$1-p$

h) Vis at middelværdien af X nu giver følgende, udtrykt ved k og indskuddet b:

$$E(X) = \left(1-\frac{1}{k}\right)\cdot b$$

På engelsk betegnes $\tilde{p}$ med udtrykket *implied probability* (underforstået sandsynlighed). Lad os nu antage, at en bookmaker er kommet frem til følgende teoretiske sandsynligheder for resultaterne af en fodboldkamp:

1 (hjemmesejr): p_1, x (uafgjort): p_2, 2 (udesejr): p_3

Det er klart, at disse virkelige sandsynligheder tilsammen giver 1: $p_1 + p_2 + p_3 = 1$. De tilhørende teoretiske odds er da givet ved:

$$odds_1 = \frac{1}{p_1}, \quad odds_2 = \frac{1}{p_2}, \quad odds_3 = \frac{1}{p_3}$$

Senere justeres disse teoretiske odds på samme måde, som ovenfor ved at vi fremskriver sandsynlighederne $\tilde{p}_1 = k \cdot p_1$, $\tilde{p}_2 = k \cdot p_2$, $\tilde{p}_3 = k \cdot p_3$, som giver anledning til nye odds:

$$\widetilde{odds}_1 = \frac{1}{\tilde{p}_1}, \quad \widetilde{odds}_2 = \frac{1}{\tilde{p}_2}, \quad \widetilde{odds}_3 = \frac{1}{\tilde{p}_3}$$

i) Vis, at der gælder: $\dfrac{1}{odds_1} + \dfrac{1}{odds_2} + \dfrac{1}{odds_3} = 1$.

j) Vis, at der gælder: $\dfrac{1}{\widetilde{odds}_1} + \dfrac{1}{\widetilde{odds}_2} + \dfrac{1}{\widetilde{odds}_3} = k$

k) Antag, at $k = 1{,}04$. Bestem bookmakerens profit i procent under antagelse af ovenstående model. *Hjælp*: Anvend formlen for middelværdien fra spørgsmål h).

l) Gå ind på nogle betting sites, fx *Oddset* under nogle fodboldkampe og aflæs odds og udregn summen af de reciprokke odds, ligesom i j). Hvad får du k til? Hvor stor er bookmakerens profit i procent?

m) Odds på en hest i et hestevæddeløb er 5.25. Omregn disse europæiske odds til britiske (fractional) odds og amerikanske (Moneyline) odds efter formlerne i tabellen nedenfor. Du kan kontrollere dit resultat på en af de mange odds convertere, som finder på Internettet, fx: https://mybettingsites.co.uk/bet-calculator/odds-converter/

Forskellige typer odds (p er den underforståede sandsynlighed)			
1	Europæisk	Decimal Odds	$odds = \frac{1}{p}$
2	Britisk	Fractional Odds	$odds = \frac{1-p}{p}$
3	Amerikansk	Moneyline Odds	$odds = \begin{cases} \frac{1-p}{p}\cdot 100 & \text{for } p \le 0{,}5 \\ -\frac{p}{1-p}\cdot 100 & \text{for } p > 0{,}5 \end{cases}$

1: Angiver den totale gevinst (inklusiv indskud) i forhold til indskuddet
2: Angiver profitten i forhold til indskuddet. Angives normalt som brøker mellem hele tal.
3: Hvis positivt: Den profit man vil få ved et indskud på \$100. Hvis negativt: Det beløb, man behøver at sætte på spil for at få en profit på \$100.

3.2 Binomialfordelingens sandsynligheder

Opgave 301

Der kastes med en terning 40 gange. I denne opgave må du kun benytte formlen for punktsandsynligheder i sætning 3.2 til at løse nedenstående spørgsmål.

a) Bestem sandsynligheden for at få netop 5 toere.
b) Bestem sandsynligheden for, at der højst er 2 toere.
c) Bestem sandsynligheden for, at der er mindst 2 toere.

Opgave 302

En ægte mønt kastes 23 gange. I denne opgave må du kun benytte formlen for punktsandsynligheder i sætning 3.2 til at løse nedenstående spørgsmål.

a) Bestem sandsynligheden for at mønten viser krone i netop 10 af kastene.
b) Hvad er sandsynligheden for at mønten viser krone mindst 10 og højst 12 gange?

Opgave 303

Løs nedenstående to uafhængige opgaver. I denne opgave må du kun benytte formlen for punktsandsynligheder i sætning 3.2 til at løse nedenstående spørgsmål.

a) Bestem sandsynligheden for at få netop 4 seksere ved 20 slag med en terning.
b) Hvad er sandsynligheden for at få mindst 11 rigtige på en tipskupon, hvis man benytter "sypigetips", dvs. man sætter krydset tilfældigt? Det oplyses, at der er i alt 13 kampe på kuponen. *Hjælp*: Hvad er basiseksperimentet og basissandsynligheden?

Opgave 304 (Farveblindhed)

Man regner med, at 5% af drengene i Danmark er farveblinde. I det følgende betragter vi en klasse, hvori der er 25 drenge. Vi interesserer os for antallet af farveblinde drenge.

a) Redegør for, hvorfor binomialfordelingen kan bruges til at beskrive situationen, ved at forklare hvad basiseksperimentet er, om der er uafhængighed, hvad succes er, etc.
b) Hvad er sandsynligheden for, at der netop er 1 farveblind dreng i klassen?
c) Hvad er sandsynligheden for, at der i klassen findes mindst én dreng, som er farveblind?

På hele skolen er der 300 drenge.

d) Hvad er sandsynligheden for, at der er mere end 25 farveblinde drenge på skolen?
e) Hvad er det gennemsnitlige antal farveblinde drenge på en skole af den nævnte størrelse?

Opgave 305

En stokastisk variabel X er binomialfordelt med antalsparameter 30 og sandsynlighedsparameter 0,18.

a) Bestem $P(X = 6)$.
b) Bestem $P(X \leq 7)$.
c) Bestem $P(4 \leq X \leq 8)$.
d) Bestem $P(X \geq 4)$.

Opgave 306

En stokastisk variabel X er binomialfordelt med antalsparameter 45 og sandsynlighedsparameter 0,37.

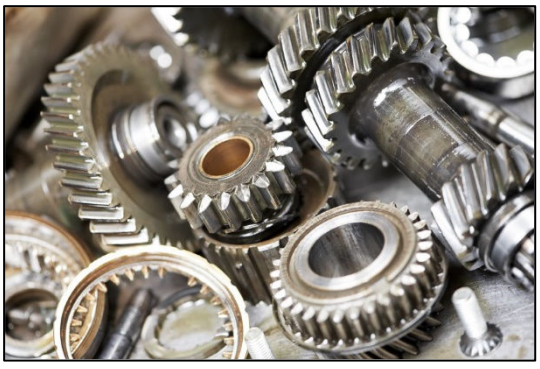
a) Bestem $P(X = 20)$.
b) Bestem $P(X \leq 22)$.
c) Bestem $P(12 \leq X \leq 18)$.
d) Bestem middelværdi, varians og spredning for den stokastiske variable X.
e) Hvad er den mest sandsynlige værdi for X?

Opgave 307 (Rygning i gymnasiet)

Ifølge Kræftens Bekæmpelse var der i 2019 i gymnasiet 9% af pigerne, som var *daglige rygere* og andre 16%, som var *lejlighedsvise rygere*. Der udtages på tilfældig måde en stikprøve på 640 piger i blandt danske gymnasieelever i 2019.

a) Bestem sandsynligheden for at maksimalt 55 af pigerne er daglige rygere.
b) Bestem sandsynligheden for at mindst 100 af pigerne er lejlighedsvise rygere.
c) Hvor stor en del af pigerne i en stikprøve af nævnte størrelse vil i gennemsnit enten være daglige ryger eller lejlighedsvise rygere?

Opgave 308 (Komponent til bil)



På en fabrik produceres en bestemt komponent til en bil. Det vides erfaringsmæssigt, at 12% af komponenterne har en defekt. Der udtages på tilfældig vis 50 komponenter.

a) Hvad er sandsynligheden for, at netop 7 komponenter har defekten?
b) Hvad er sandsynligheden for i stikprøven at få mindst 6 og højst 8 defekte komponenter?
c) Hvad er sandsynligheden for at mindst 4 komponenter er defekte?
d) Hvad er det gennemsnitlige antal defekte komponenter i en stikprøve af den nævnte størrelse?
e) Hvor stor er spredningen?

Opgave 309 (Blodtyper)

I den danske befolkning er der følgende fordeling af blodtyper:

	Rhesus positiv	Rhesus negativ
A	37%	7%
B	8%	2%
0	35%	6%
B	4%	1%

Der udtages en tilfældig gruppe på 50 personer fra den danske befolkning.

a) Hvad er sandsynligheden for, at der netop er fem B Rhesus positive i gruppen?
b) Hvor mange B Rhesus positive vil der i gennemsnit være i en gruppe på 50 personer?
c) Bestem sandsynligheden for, at der er mindst seks med blodtype 0 Rhesus negativ.

Opgave 310 (Tulipanløg)

Af erfaring vides det, at 83% af tulipanløgene for en bestemt type tulipaner spirer. Et gartneri anskaffer 200 tulipanløg.

a) Redegør for, hvorfor binomialfordelingen kan bruges til at udregne sandsynlighederne for rækken af opgaver nedenfor.
b) Bestem sandsynligheden for at netop 165 tulipanløg spirer.
c) Bestem sandsynligheden for at maksimalt 165 af tulipanløgene spirer.
d) Bestem sandsynligheden for at der er mellem 160 og 175 af tulipanløgene, som spirer, begge værdier inklusive.
e) Bestem sandsynligheden for at mindst 175 af tulipanløgene spirer.
f) Hvor mange tulipanløg vil i gennemsnit spire?
g) Bestem variansen og spredningen i antallet af spirene tulipanløg
h) Hvad er det mest sandsynlige antal spirende tulipanløg?
i) Tegn et pindediagram over sandsynlighedsfordelingen.
j) Afgør om binomialfordelingen er venstreskæv, højreskæv eller central.
k) Afgør om 150 spirende tulipanløg er et normalt eller exceptionelt udfald eller ingen af delene (se definition 2.26).

Opgave 311 (Multiple Choice)

I en multiple-choice prøve er der 25 spørgsmål med hver 5 mulige svar. Antag, at du ikke ved et klap om emnet og svarer helt tilfældigt.

a) Hvad er da sandsynligheden for at få mindst 10 rigtige i prøven?
b) Hvad er middeltallet for antal rigtige svar?

Opgave 312 (Lotteri)

I et lotteri kan der trækkes lodder. Der er 5% chance for at få en gevinst på et givet lod.

a) Hvor stor er sandsynligheden for ved 14 trækninger at få mindst 2 gevinstlodder?
b) Hvor mange lodder skal man mindst trække, før sandsynligheden for at der er gevinst på mindst ét af dem overstiger 90%.

Opgave 313

Givet en binomialfordelt stokastisk variabel X med antalsparameter 17 og sandsynlighedsparameter 0,75.

a) Lav et pindediagram over sandsynlighedsfordelingen for X.
b) Er der tale om en venstreskæv, højreskæv eller central binomialfordeling, jf. definition 3.8?

Opgave 314 (Fodboldkort)

Rasmus samler på kort med danske og internationale fodboldspillere. 15% af kortene har en dansk spiller afbildet. Rasmus køber 60 kort. Man interesserer sig for, hvor mange af kortene, som har danske spillere afbildet.

a) Redegør for, hvorfor binomialfordelingen kan bruges til at beskrive situationen, ved at forklare hvad basiseksperimentet er, om der er uafhængighed, hvad succes er, hvad den stokastiske variabel X skal stå for, etc.
b) Hvad er sandsynligheden for, at netop 7 af kortene er med danske spillere?
c) Hvad er sandsynligheden for, at der er mindst 10 kort med danske spillere?
d) Hvad er sandsynligheden for, at der er enten 5 kort med danske spillere eller 12 kort med danske spillere?
e) Hvor mange kort vil i gennemsnit indeholde danske spillere, når man køber 60 kort?
f) Lav et pindediagram over sandsynlighedsfordelingen for den stokastiske variabel X.

Rasmus har fået oplyst, at 1,5% af kortene har Christian Eriksen som motiv.

g) Hvor mange kort skal Rasmus mindst købe, hvis han skal være mindst 80% sikker på at få et kort med Christian Eriksen?

Opgave 315 (Kattemad)

På en fabrik fylder maskiner automatisk kattemad i poser. Det medfører små variationer i nettovægten. Selvom man har justeret maskinerne, så de gennemsnitligt fylder lidt mere i poserne, end der står i deklarationen, viser tests, at der er 4% af poserne, hvor nettovægten er under det deklarerede. Der udtrækkes på tilfældig måde 200 poser.

a) Hvad er sandsynligheden for, at der højst er 5 poser, som vejer for lidt?
b) Bestem sandsynligheden for, at der er mindst 7 og højst 10 poser, som vejer for lidt.
c) Bestem middelværdien og spredningen for den stokastiske variabel X, som angiver antallet af poser med for lav vægt.
d) Hvad er de *exceptionelle* værdier for X, og hvad er sandsynligheden for, at de forekommer (se definition 2.26).

Opgave 316 (Kombi-opgave)

Der kastes med to terninger. Vi vælger at kalde det en *succes*, hvis summen af terningernes visning *enten* er 5 *eller* den ene terning viser en firer.

a) Redegør for, at sandsynligheden for succes er $\frac{13}{36}$.

Der kastes 12 gange med to terninger.

b) Hvad er sandsynligheden for at få succes i netop 5 af kastene?
c) Hvad er sandsynligheden for at få mindst 3 succeser i de 12 kast?

Opgave 317

Rolf er basketball spiller og er generelt set i stand til at ramme kurven i 80% af kastene. I det følgende betyder L et kast der lykkes og M et kast, der mislykkes.

a) Bestem sandsynligheden for følgende serie i 10 kast: $MLMMLLLLML$.
b) Hvad er sandsynligheden for at Rolf rammer kurven i 8 ud af 10 kast.

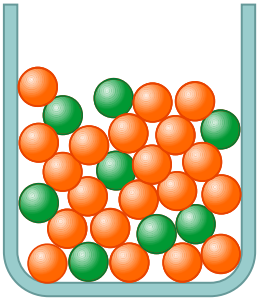
Opgave 318

En stokastisk variabel X er binomialfordelt med antalsparameter 85 og sandsynlighedsparameter 0,22.

a) Bestem $P(X=15)$ b) Bestem $P(X \le 20)$ c) Bestem $P(X > 25)$.

d) Hvad er den mest sandsynlige værdi for X?

Opgave 319 (Med og uden tilbagelægning)



Som bekendt kræver anvendelsen af binomialfordelingen, at der ved udførelse af hvert basiseksperiment er den samme sandsynlighed. Man omtaler ofte problematikken med, at det skal være *med tilbagelægning* således, at udgangspunktet ved udførelsen af næste basiseksperiment er uændret. I eksempel 3.9 med folketingsvalg var dette ikke opfyldt, men vi kunne se bort det, da sandsynligheden ændrede sig yderst lidt grundet den lille stikprøve taget fra den store population. I denne opgave skal vi se en situation, hvor det har en klar betydning for sandsynlighederne, om der er tale om med tilbagelægning eller uden tilbagelægning. Der er givet en krukke indeholdende 20 orange kugler og 8 grønne kugler.

a) Der trækkes nu 10 kugler *med* tilbagelægning fra krukken. Hvad er sandsynligheden for at få 5 grønne kugler?

b) Der trækkes nu 10 kugler *uden* tilbagelægning fra krukken. Hvad er sandsynligheden for at få 5 grønne kugler? *Hjælp*: Benyt teknikken fra kapitel 2 med at kigge på antal gunstige og antal mulige udfald.

Opgave 320 (Folketingsvalg)

Op til folketingsvalg foretages som bekendt mange valgprognoser. Typisk udspørges mellem 1000 og 1500 personer om, hvilket parti de vil stemme på. I dette eksempel skal vi forsøge at få en fornemmelse for, hvor usikre sådanne stikprøver i virkeligheden er. Lad os antage, at et parti i virkeligheden på landsplan har en tilslutning på præcist 20%. Vi antager nu, at der foretages en valgprognose på baggrund af en stikprøve bestående af 1200 tilfældigt udvalgte personer.

a) Hvad er sandsynligheden for, at stikprøven på de 1200 vil give et resultat, som afviger med mere end 2 procentpoint fra partiets rigtige tilslutning?

   Hjælp: Fra 18% til 22%. Hvad svarer det til i antal stemmer?

Opgave 321

En stokastisk variabel X kan antage værdierne 0, 1, … , 10. Følgende sandsynligheder er oplyst: $P(X \geq 4) = 0{,}60$, $P(X \leq 7) = 0{,}70$ og $P(5 \leq X \leq 7) = 0{,}12$. Bestem $P(X = 4)$.
Hjælp: Tegn en linje med de mulige værdier af X og indram de involverede hændelser.

3.3 Approksimation med normalfordelingen

Opgave 322 (Tæthedsfunktionen)

Tegn graferne for tæthedsfunktionerne for følgende normalfordelinger i samme koordinatsystem: $f_{0,1}$, $f_{0,0.5}$ samt $f_{4,2}$ i samme koordinatsystem.

Hjælp: For at være mere økonomisk, kan du eventuelt vælge at definere en funktion af tre variable $f(\mu,\sigma,x)$, hvorefter de tegner graferne for $f(1,0,x)$, $f(0,0.5,x)$ og $f(4,2,x)$. Det kan også være, at du har en indbygget funktion i dit CAS-værktøj til formålet.

Opgave 323 (Fordelingsfunktionen)

Tegn graferne for fordelingsfunktionerne for følgende normalfordelinger i samme koordinatsystem: $F_{0,1}$, $F_{0,0.5}$ samt $F_{4,2}$ i samme koordinatsystem.

Hjælp: Du skal her bruge nogle indbyggede funktioner, som må være til stede i dit CAS-værktøj. At definere funktionerne via integralerne (6) er ofte for tungt.

Opgave 324 (Værnepligtiges højde)

Ved at analysere højderne af 13427 værnepligtige i Danmark fra andet halvår af 2006 kan man konkludere, at soldaternes højde med meget stor nøjagtighed følger en normalfordeling med middelværdi 180,1 cm og spredning 6,81 cm. Med disse oplysninger skal nedenstående spørgsmål besvares. *Hjælp*: Se teknikken i eksempel 3.11.

a) Hvor mange procent af de værnepligtige mænd har en højde på under 170 cm?
b) Hvor mange procent har en højde på mellem 175 cm og 180 cm?
c) Hvor mange procent har en højde på mindst 200 cm?

Opgave 325

Påvis, at arealet under grafen for $f_{\mu,\sigma}$ fra $-\infty$ til ∞ virkelig er 1, når $\sigma > 0$.

Opgave 326

Givet en stokastisk variabel $X \sim N(50,3)$.

a) Bestem sandsynligheden $P(X \le 48)$ ved at integrere tæthedsfunktionen givet ved (5) i afsnit 3.3.

b) Tegn grafen for tæthedsfunktionen og skraver om muligt området under grafen bestemt af $X \le 48$.

c) Udregn desuden sandsynligheden $P(X \le 48)$ ved at benytte fordelingsfunktionen, gerne med en indbygget funktion fra CAS-værktøjet.

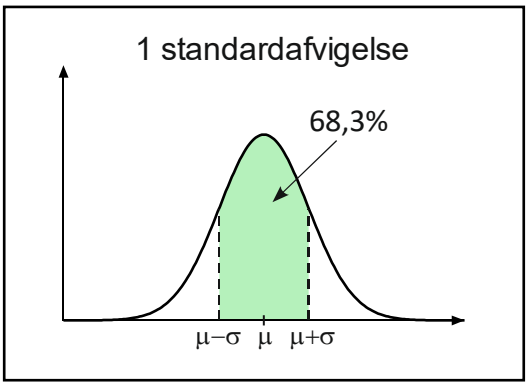
d) Tegn grafen for fordelingsfunktionen og marker løsningen til $P(X \le 48)$.

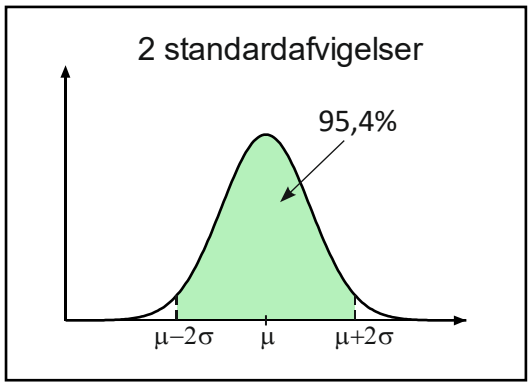
Opgave 327 (En egenskab for normalfordelingen)

Der gælder nogle smukke egenskaber for en normalfordelt stokastisk variabel X, nemlig at sandsynligheden for, at x ligger indenfor én spredning, også kaldet *standardafvigelse*, fra middelværdien altid giver 68,3%. Tilsvarende gælder, at sandsynligheden for, at x ligger maksimalt 2 spredninger fra middelværdien, altid er lig med 95,4%. Opskrevet på matematisk form er det:

$$P(\mu - \sigma \le X \le \mu + \sigma) = 68{,}3\%$$
$$P(\mu - 2\sigma \le X \le \mu + 2\sigma) = 95{,}4\%$$

Denne egenskab benyttes i mange grene af statistikken.





a) Vis med dit CAS-værktøj, at disse påstande er rigtige. *Hjælp*: Udnyt enten (5) eller (7) i afsnit 3.3.

Man kalder værdier af X, som er højst 2 spredninger fra middelværdien for *normale*, mens værdier, som ligger mere end 3 spredninger fra middelværdien, for *exceptionelle*.

b) Hvad er sandsynligheden for, at X antager en exceptionel værdi?

Opgave 328 (Approksimere binomialfordeling med normalfordeling)

Betragt en binomialfordelt stokastisk variabel X med antalsparameter 25 og sandsynlighedsparameter 0,36, dvs. $X \sim b(25, 0.36)$.

a) Bestem middelværdi og spredning for X.

Vi ønsker at tilnærme binomialfordelingen med en normalfordeling. Det er i den forbindelse en tommelfingerregel, at følgende betingelser skal være opfyldt, før approksimationen er tilstrækkelig god:

$$n > 9 \cdot \left(\frac{p}{1-p} \right) \text{ og } n > 9 \cdot \left(\frac{1-p}{p} \right)$$

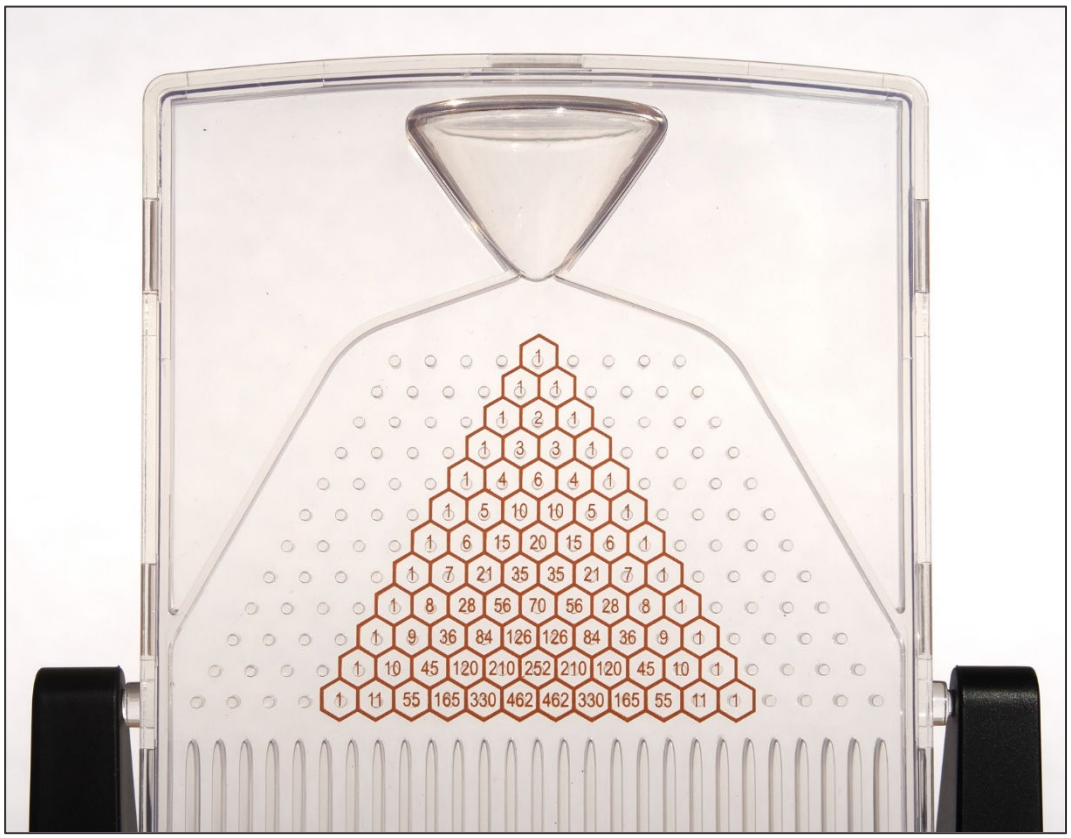
b) Vis, at begge betingelser er opfyldt.

c) Forsøg at lave en kombineret graf, bestående af pindediagrammet for binomialfordelingen og grafen for tæthedsfunktionen for den normalfordeling, som har samme middelværdi og spredning som beregnet i a). Ser tilnærmelsen god ud?

Opgave 329 (Galtons bræt og Pascals trekant - projekt)

I denne opgave skal vi kigge på et instrument kaldet et *Galton Board*, opkaldt efter den berømte britiske statistiker Sir Francis Galton (1822-1911).

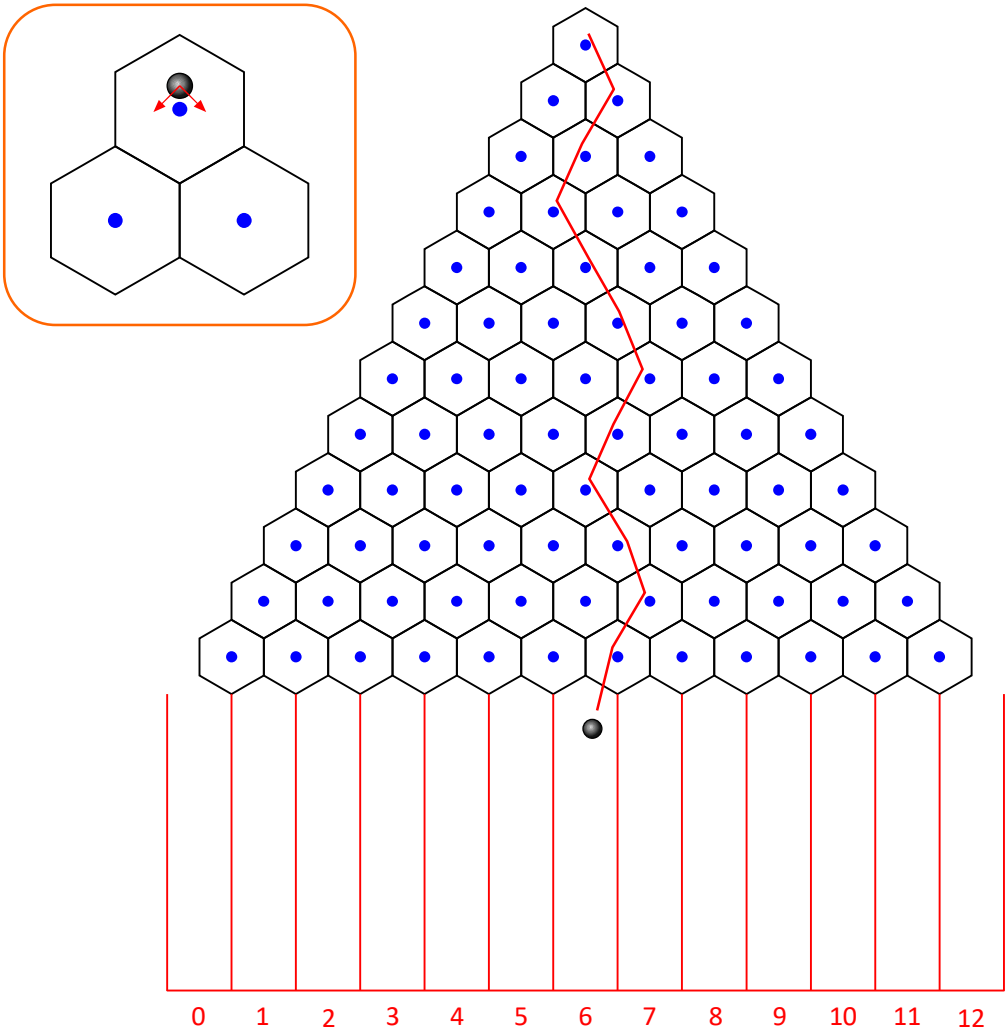
Princippet i et Galton Board eller Galton bræt er, at der er en mængde kugler, som falder lodret ned til et netværk af pinde. Når en kugle rammer en pind, kan den enten gå til højre eller til venstre. Når kuglen er kommet forbi pindene, falder den ned i en beholder umiddelbart under. Når en stor mængde af kugler (i tilfældet nedenfor 3000 kugler) løber ned gennem brættet, vil man på grund af den indbyggede tilfældighed og de store tals lov ende op med et slags histogram, som på næsten magisk vis ligner en klokkekurve. Man vil se, at histogrammet meget pænt tilnærmes af den indtegnede normalfordelingskurve. Kloge folk har fascineret betegnet det som en form for *orden i kaosset*!



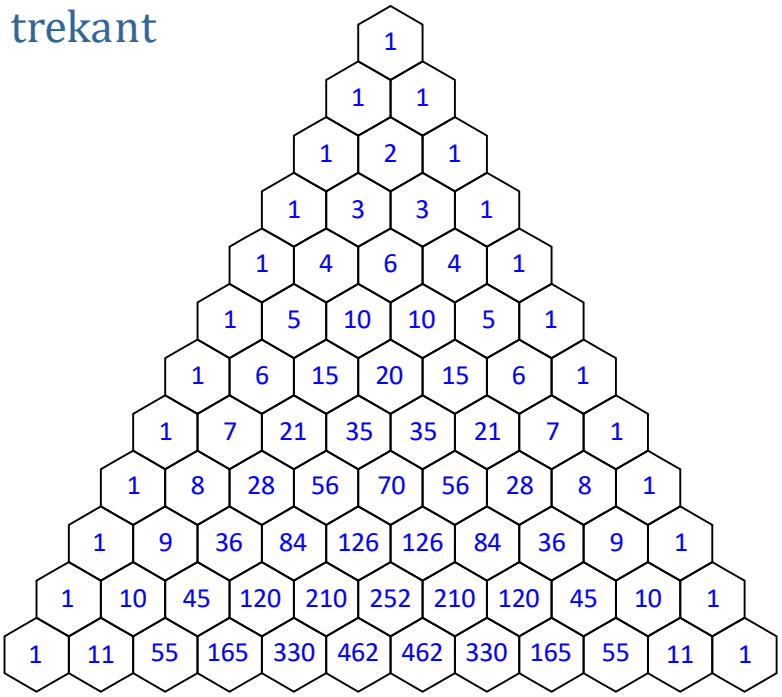
a) Find videoer på YouTube, hvor et Galton bræt (Galton Board), undertiden også kaldet *quincunx*, demonstreres og forklares.

På figuren øverst på næste side er princippet i et idealiseret Galton bræt vist. I den orange ramme øverst til venstre på figuren er vist, at når en kugle kommer til en pind, så kan den gå to veje: enten mod venstre eller mod højre, og det med lige sandsynlighed. Resten af figuren viser en mulig *vej* for en kugle ned igennem brættet. Hvis vi med V mener venstre og med H mener højre, så kan den afbillede kugles vej skrives HVVHHHVVHHVV. Under sektionen af pinde er anbragt nogle beholdere til at samle kuglerne op i. Vi har nummereret dem 0, 1, …, 12. I det pågældende tilfælde endte kuglen i beholder nr. 6.

I det følgende vil vi antage, at der er n lag af pinde, som kuglerne skal igennem, og at kuglerne derved ender i beholdere nummereret 0, 1, …, n.



Pascals trekant



b) Redegør for, hvorfor antallet af *veje* til den *k*'te beholder er $K_{n,k}$.

Hjælp: Argumenter for, at hvis kuglen går mod højre i *k* tilfælde, så vil kuglen ende i beholder nr. *k*, uanset rækkefølgen af højre-venstre. Benyt derefter sætning 1.11.

c) Vis, at sandsynligheden for, at en kugle ender i beholder nr. *k*, er $K_{n,k} \cdot \left(\frac{1}{2}\right)^n$.

Hjælp: Argumenter for, at vi kan bruge binomialfordelingen med antalsparameter *n* og sandsynlighedsparameter $\frac{1}{2}$. Hvad er basiseksperimentet? Hvad vil det sige, at et basiseksperiment lykkes/mislykkes?

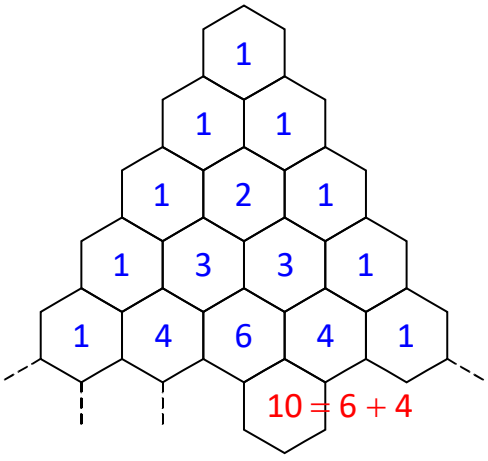
d) Udregn for tilfældet $n = 12$ en liste over sandsynlighederne fra c) for $k = 0, 1, \ldots, n$.

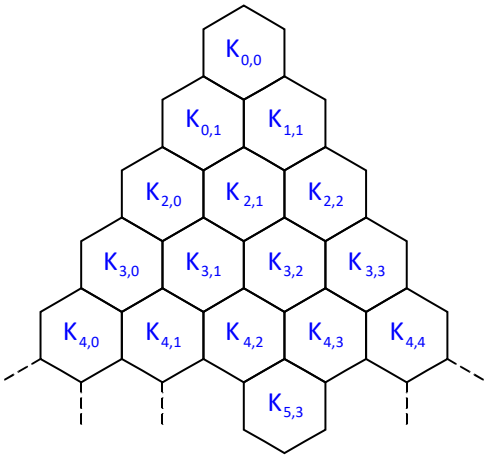
e) Udregn middelværdi μ og spredning σ for den binomialfordeling, som er omtalt i hjælpen til spørgsmål c). Tegn derefter for tilfældet $n = 12$ pindediagrammet for binomialfordelingen og tæthedsfunktionen for normalfordelingen med parametre μ og σ i samme koordinatsystem. Er normalfordelingen en god tilnærmelse?

Som det ses på de forrige sider, er den aktuelle udformning af Galtons bræt ikke helt som det ideelle. Der er nogle ekstra pinde i hver side, ligesom der er nogle ekstra beholdere udover det, der svarer til antal rækker af pinde + 1. Det viser sig at fungere bedst på denne måde. Vi ser også den fine klokkekurve her.

f) Prøv at give nogle bud på, hvorfor det kan være kompliceret at få den fysiske udgave af brættet til at opføre sig præcist som den ideelle og teoretiske model. Hvad kan der ske med kuglerne i praksis …?

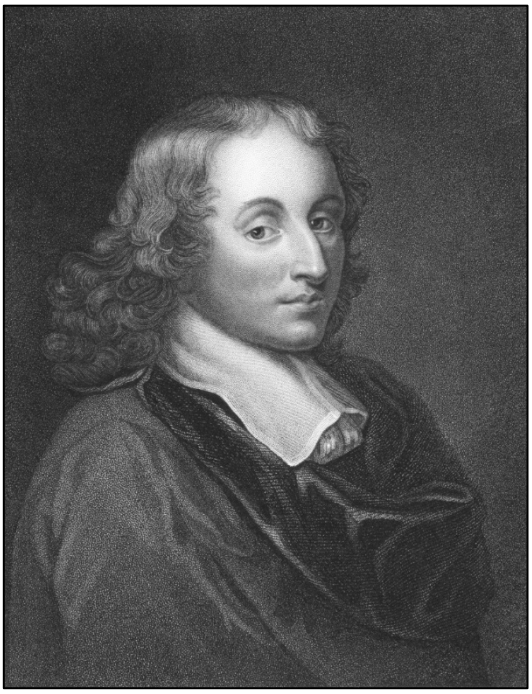
Vi skal nu betragte en anden facet af ovenstående problematik, nemlig *Pascals trekant*, som er afbildet nederst på forrige side for tilfældet $n = 12$. Pointen er, at man lægger sammen fra oven, som markeret med rødt herunder: Den nye værdi nedenfor er lig med summen af de to umiddelbart ovenfor. Vi ser desuden til højre, at alle værdierne svarer til de velkendte kombinationer $K_{n,k}$.





f) Kig på Pascals trekant et par sider tilbage. Forsøg at finde mønstre i talværdierne. Hvilke sammenhænge synes der at gælde? Hvis du kan argumentere for dem, er det fantastisk!

Faktisk er der et væld af smukke sammenhænge for tallene i Pascals trekant. Noget så specielt som *Fibonacci-tallene* dukker endda op. Ikke mere om det her. Til sidst skal det lige nævnes, at både Pascal og Galton spillede centrale roller i sandsynlighedsregningen og statistikken. Allerede som børn markerede de sig som de genier, de var.



Blaise Pascal (1623-1662)

Sir Francis Galton (1822-1911)

Opgave 330 (Meningsmålinger – spredning ved kendt sandsynlighedsparameter)

I denne opgave skal vi se på den betydning stikprøvestørrelsen har hvad angår usikkerheden i forbindelse med spørgeundersøgelser, hvor der kun er to mulige svar. Det kunne være en prognose for et valg, hvor befolkningen skal svare ja eller nej til et spørgsmål eller om befolkningen stemmer på et givet parti eller ej. Det er oplagt, at der her er tale om en approksimativ binomialfordeling, så længe stikprøvestørrelsen er meget mindre end populationens størrelse. Det får os til at spørge, hvor meget vi kan regne med resultatet af en sådan stikprøve, og hvilken betydning stikprøvens størrelse n har? Vi vil antage, at vi kender sandsynlighedsparameteren p. Det gør man godt nok normalt ikke, da det jo netop er stikprøven, som skal kaste lys over det, men antagelsen vil sætte os i stand til at få et indblik i den betydning antallet af udspurgte har. Lad i det følgende X angive den binomialfordelte stokastiske variabel, som angiver antal gange basiseksperimentet lykkes, for eksempel, hvor mange, som stemte ja. Ifølge sætning 3.6 er middelværdien og spredningen for X givet ved henholdsvis $\mu = n \cdot p$ og $\sigma = \sqrt{n \cdot p \cdot (1-p)}$.

Vi ved, at når n er tilstrækkelig stor (se for eksempel tommelfingerreglerne i opgave 328), så er normalfordelingen med samme middelværdi og spredning en god approksimation til binomialfordelingen. Det betyder, at vi kan gøre brug af den smukke egenskab, som gælder for en normalfordelt stokastisk variabel (se opgave 327):

$$P(\mu-2\sigma\le X\le\mu+2\sigma) \;=\; 95{,}4\%$$

Vi kan foretage nogle omskrivninger:

$$\begin{aligned} P(\mu-2\sigma\le X\le\mu+2\sigma) \;&=\; P\left(\frac{\mu-2\sigma}{n}\le\frac{X}{n}\le\frac{\mu+2\sigma}{n}\right)\\ &=\; P\left(\frac{n\cdot p-2\cdot\sqrt{n\cdot p\cdot(1-p)}}{n}\;\le\;\frac{X}{n}\;\le\;\frac{n\cdot p+2\cdot\sqrt{n\cdot p\cdot(1-p)}}{n}\right)\\ &=\; P\left(p-2\cdot\sqrt{\frac{p\cdot(1-p)}{n}}\;\le\;\frac{X}{n}\;\le\;p-2\cdot\sqrt{\frac{p\cdot(1-p)}{n}}\right) \end{aligned}$$

Først lighedstegn: Vi dividerede alle sider i den dobbelte ulighed med den positive konstant n. Andet lighedstegn: Vi indsatte de kendte værdier for middelværdi og spredning. Tredje lighedstegn: Vi reducerede udtrykkene. Alt i alt får vi:

$$\text{(A)}\qquad P\left(p-\Delta p\;\le\;\frac{X}{n}\;\le\;p+\Delta p\right)\;\approx\;95{,}4\%\quad\text{hvor}\quad\Delta p\;=\;2\cdot\sqrt{\frac{p\cdot(1-p)}{n}}$$

Husk at n er stikprøvestørrelsen. Ovenfor kan X/n dermed tolkes som den brøkdel af de adspurgte, som stemte ja (eller hvad det nu var). Dette tal behøver ikke være lig med p, for det er en stikprøve af populationen, vi har udspurgt. Men hvis stikprøven er foretaget tilfældigt og repræsentativt, kan vi altså regne med, at ca. 95% af sådanne stikprøver vil give et resultatet p plus minus en usikkerhed givet ved det sidste udtryk i (A).

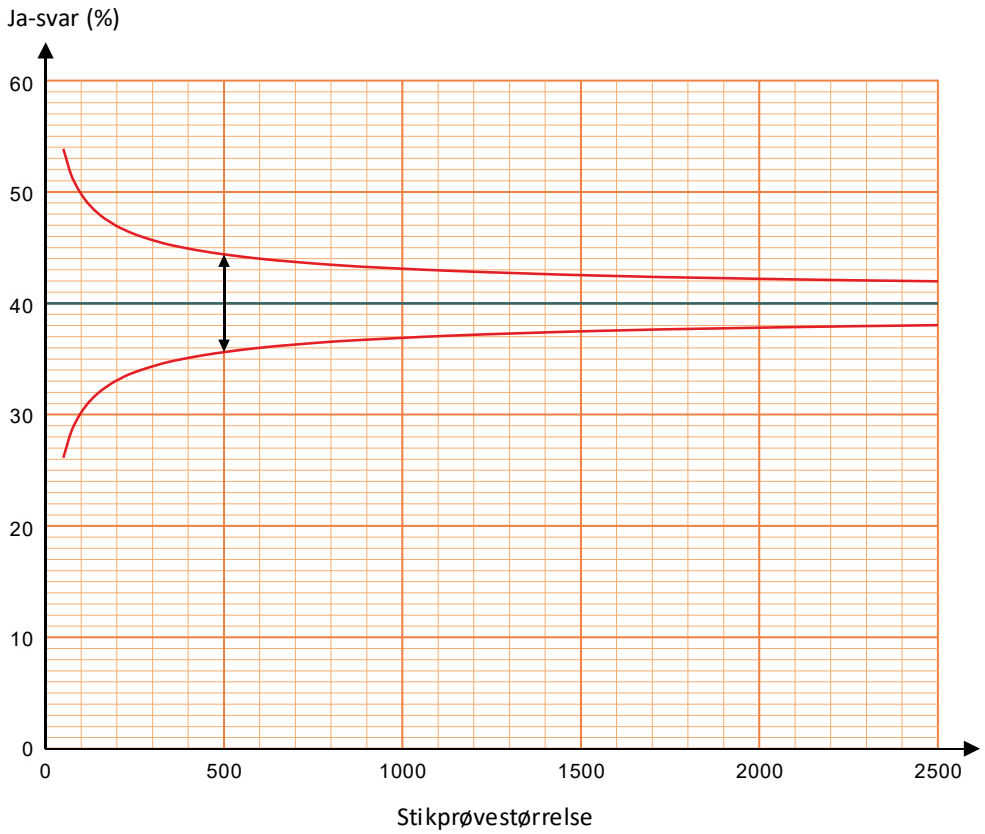
Lad os se på et eksempel: Lad os forestille os, at der i virkeligheden i populationen er 40%, som ville sige ja, og vi udspurgte 500 personer tilfældigt om, hvad de ville stemme. Usikkerheden bliver i dette tilfælde:

$$\Delta p\;=\;2\cdot\sqrt{\frac{p\cdot(1-p)}{n}}\;=\;2\cdot\sqrt{\frac{0{,}40\cdot(1-0{,}40)}{500}}\;=\;0{,}0438$$

hvorved

$$\begin{aligned} p-\Delta p\;&=\;40\%-4{,}4\%\;=\;35{,}6\%\\ p+\Delta p\;&=\;40\%+4{,}4\%\;=\;44{,}4\% \end{aligned}$$

I ca. 95% af sådanne stikprøver med 500 udspurgte ville vi altså enden op med et resultat, som ligger mellem 35,6% og 44,4%. Et ikke helt lille spænd! Situationen er illustreret på figuren på næste side. Her er spændet markeret med en sort dobbeltpil. Spørgsmålet er nu, hvad der sker, hvis vi udspørger færre eller flere end 500 personer. Ja de to kurver angiver grænserne for det interval, som ca. 95% af stikprøverne vil lande indenfor. Man ser, at spændet bliver mindre jo flere der spørges. Man kunne så indvende, at det bare gælder om at spørge så mange som muligt, for så får man et mere sikkert resultat. Her skal man dog huske på, at sådanne prognoser kan være ganske dyre at foretage.



a) Beregn usikkerheden Δp og spændet, når der udspørges 1000 personer.

b) Hvor mange skal man spørge for at usikkerheden er nede på 2 procentpoints?

c) Kig på udtrykket for Δp i (A). Hvad sker der med usikkerheden, når man firdobler stikprøvestørrelsen?

d) Forsøg selv at plotte yderkurverne for tilfældet $p = 0{,}15$. Hvad registrerer du om spændet i procentpoint, når p er lavere?

Hjælp: Plot grafen for funktionen

$$p_{\text{øvre}}(t) = 100 \cdot \left(p + 2 \cdot \sqrt{\frac{p \cdot (1-p)}{t}} \right)$$

i intervallet givet ved $0 \le t \le 2500$. Tilsvarende for $p_{nedre}(t)$.

Opgave 331 (Hvor nøjagtig er approksimationen?)

Prøv med dit værktøjsprogram at undersøge, hvor god en approksimation normalfordelingen er til binomialfordelingen. Du kan for eksempel undersøge tilfældet $n = 50$, $p = 0{,}25$. Udregn først middelværdi og spredning for normalfordelingen via formlerne i sætning 3.6, altså $\mu = n \cdot p$ og $\sigma = \sqrt{n \cdot p \cdot (1-p)}$. Bestem derefter de kumulerede sandsynligheder *bincdf* $(n, p, 20)$ og *normalcdf* $(\mu, \sigma, 20)$. Hvor stor er fejlen? Benyt derefter *kontinuitetskorrektionen* nævnt i (9) side 44: Dvs. man lægger 0,5 til: *normalcdf* $(\mu, \sigma, 20.5)$. Hvor god er approksimationen nu? NB! Det skal nævnes, at når n bliver større og større, bliver betydningen af at benytte kontinuitetskorrektionen mindre og mindre!

Opgave 332 (Vægten af nyfødte grise)

På hjemmesiden svineproduktion.dk er det nævnt, at vægten af nyfødte grise er normalfordelt med en middelværdi på 1,5 kg. Desuden er det oplyst, at 15% af grisene har en fødselsvægt på under 1 kg.

a) Benyt oplysningerne til at vise, at $\sigma = 0{,}4824$. *Hjælp*: Der er flere måder at løse dette på, for eksempel ved at opstille en ligning, som involverer integralet (6) side 41, hvor σ er den ubekendte, eller man kan bruge (8) samt den inverse funktion til fordelingsfunktionen for standardnormalfordelingen, dvs. Φ^{-1}.
b) Med den fundne værdi for spredningen i a) skal du bestemme sandsynligheden for at der en nyfødt gris vejer mere end 2,5 kg.
c) Hvad er sandsynligheden for at grisen vejer mellem 1,2 kg og 1,8 kg?

Opgave 333

En stokastisk variabel X er normalfordelt med middelværdi 40 og spredning 7.

a) Bestem sandsynligheden $P(X \le 35)$.
b) Bestem sandsynligheden $P(38 \le X \le 44)$.
c) Bestem sandsynligheden $P(X \ge 50)$.

Opgave 334 (Eksperimentel støj)

Når man foretager eksperimenter, vil der ofte indtræffe tilfældige fejl i målingerne, som enten er forårsaget af ukendte og uforudsigellige ændringer i eksperimentet, eller som har rod i måleinstrumenterne. Det sidste kan for eksempel være støj i et elektronisk instrument. Disse tilfældige fejl har ofte en normalfordeling. Lad os antage, at den korrekte og ukendte middelværdi μ for en måling er 47,9 mm og at den ligeledes ukendte spredning er 2,8 mm. Hvad er så sandsynligheden for, at man i eksperimentet opnår et måleresultat på 48,8 mm eller derunder – under antagelse af en normalfordeling?

Opgave 335 (Tilnærme binomialfordelingen med en normalfordeling)

Som bekendt kan man benytte binomialfordelingen til at bestemme sandsynligheden for at få højst 14 seksere ved 100 kast med en terning. Men man kan også bruge normalfordelingen til at give et tilnærmet bud på sandsynligheden.

a) Brug binomialfordelingen til at bestemme en eksakt værdi for sandsynligheden.
b) Udregn middelværdien og spredningen for binomialfordelingen efter formlerne i sætning 3.6. Benyt derefter normalfordelingens fordelingsfunktion med disse værdier for μ og σ til at bestemme en tilnærmet værdi for sandsynligheden for højst 14 seksere i 100 kast – både med og uden kontinuitetskorrektionen (9) side 44.
c) Lav en kombineret graf, bestående af pindediagrammet for binomialfordelingen og grafen for tæthedsfunktionen for normalfordelingen. Ser approksimationen pæn ud?

4.2 Binomialtest

Opgave 401

Man er i tvivl, om en terning slår ettere med den frekvens, der forventes, dvs. 1/6. Derfor udføres et forsøg, hvor man slår 80 gange med terningen. Ved forsøget blev der registreret i alt 19 ettere.

a) Opstil en nulhypotesen og en alternative hypotese?
b) Hvor mange ettere vil man i middel få, hvis nulhypotesen er sand?
c) Afgør med en tosidet binomialtest med signifikansniveau 5%, om frekvensen af ettere fra terningen kan antages at være 1/6.

Opgave 402 (Rollespil)

Der er mistanke om, at en rollespilsterning med 12 sider ikke giver den korrekte frekvens af 12'ere. Derfor foretager man en stikprøve, hvor man kaster terningen 500 gange. Ved den lejlighed blev der slået 51 12'ere.

a) Opstil en nulhypotese og afgør ved et signifikansniveau på 5%, om rollespilsterningen mon giver den rigtige frekvens af 12'ere.
b) Samme spørgsmål, hvis stikprøven havde vist 28 12'ere.

Opgave 403 (Kvalitetskontrol)

Maibrit er kvalitetskontrollør for en importør af dagligvarer. Hun skal teste, om en stor levering af æbler lever op til leverandørens lovning om, at mindst 85% af æblerne har en vægt på minimum 150 gram. For at vurdere, om leverandøren lever op til betingelserne, udtrækker hun i alt 125 æbler fra forskellige kasser og vejer dem. Derved kom hun frem til, at der var 97 af æblerne, som overholdt minimumsvægten.

a) Opstil en nulhypotese og en alternativ hypotese. Hvorfor er det rimeligt at foretage en venstresidet test her?
b) Foretag en binomialtest med et signifikansniveau på 5% for, at give en vurdering af, om æblerne i hele det store parti lever op til lovningerne.

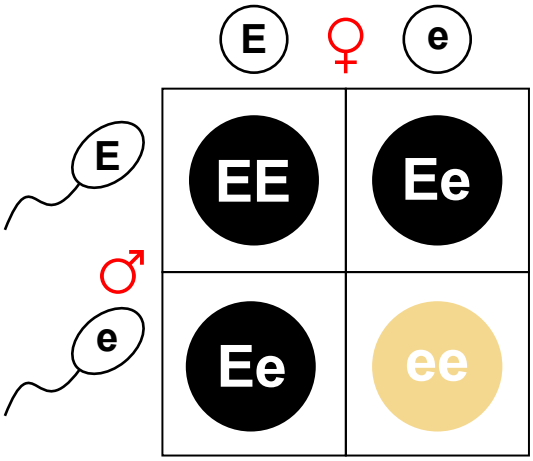
Opgave 404 (Valg)

Ved sidste kommunalvalg i en større by fik partiet H 38% af stemmerne. Der er en formodning om, at partiet er gået tilbage. Ved en vælgerundersøgelse blandt 620 indbyggere udtalte 213, at de stadig vil stemme på partiet H.

a) Opstil en nulhypotese, der kan anvendes til at teste, om tilslutningen til partiet er blevet mindre.
b) Benyt et binomialtest med et signifikansniveau på 1% til at vurdere, om tilslutningen til partiet er blevet mindre.

Opgave 405 (Genetik)

Hos hunderacen *Labrador Retriever* er der et gen, som bestemmer om pelsen bliver mørk eller gul. I det mørke tilfælde er der et andet gen, som er bestemmende for, om den mørke farve ender med at blive sort eller brun. I denne opgave ser vi imidlertid kun på førstnævnte gen. Genet har to alleler (varianter), nemlig det dominante E for mørk og det recessive e for gul. Det betyder, at hvis afkommet modtager en dominant allel fra blot en af forældrene, så ender afkommet med at blive mørk. Kun i tilfælde af, at ungen modtager en recessiv allel fra begge forældre, bliver ungen gul, som vist i krydsningsskemaet:



Det oplyses, at allelfrekvensen er $p = 0{,}70$ for E og $q = 0{,}30$ for e i Danmark.

a) Benyt allelfrekvenserne til at beregne sandsynlighederne for, at et tilfældigt valgt individ får en unge med genotyperne henholdsvis EE, Ee og ee.
b) Bestem sandsynligheden for at ungen er mørk, henholdsvis gul.

Populationen er i *Hardy-Weinberg-ligevægt*, hvis frekvenserne af de tre genotyper EE, Ee og ee i populationen er henholdsvis p^2, $2pq$ og q^2. Der udtrækkes på tilfældig vis 260 Labrador Retrievere forskellige steder i landet. Blandt dem observerede man 29 gule hunde af racen.

c) Afgør med en binomialtest med et signifikansniveau på 5%, om antallet af gule hunde er i overensstemmelse med det, som en Hardy-Weinberg-ligevægt foreskriver.
d) Læs betingelserne i Hardy-Weinberg loven nedenfor. Giv mindst ét bud på, hvad der kan få betingelserne til ikke at være opfyldt.

Hardy-Weinberg-loven: Allel-frekvenserne for et gen er konstant fra generation til generation, såfremt populationen er (uendelig) stor, parringer foregår tilfældigt, der er hverken migration, selektion eller mutationer og allelhyppighederne er ens hos hanner og hunner. Hvis disse betingelser opfyldt, kan man for diploide organismer udregne fordelingen af genotyper i populationen. Har et gen to alleler A og B med frekvenser henholdsvis p og $q = 1 - p$, så fås genotypefrekvenserne p^2, q^2 og $2pq$ for henh. AA, BB og AB (ofte omtalt som Hardy-Weinberg-ligevægten eller Hardy-Weinberg-proportionerne).

Opgave 406 (Triangeltest - kaffesmagning)

På hjemmesiden smagefter.dk er omtalt begrebet en *triangeltest*. Den benyttes, når man sætter en test op med henblik på at afgøre, om der er en lille smagsforskel på to produkter. Man opstiller tre prøver på produktet, mærket med kodenumre. To af prøverne er ens, mens den tredje er anderledes. Hver testperson skal afgøre hvilken af prøverne, som adskiller sig fra de andre. Hvis der *ikke* er nogen mærkbar smagsforskel, vil der med al sandsynlighed være ca. 1/3 af testpersonerne, som peger på den rigtige prøve, altså den, som skiller sig ud. *Er* der derimod en mærkbar smagsforskel vil der sandsynligvis være signifikant flere end 1/3 af testpersonerne, som peger på den prøve, som skiller sig ud.

I en konkret smagsprøvning skulle det vurderes, om der er smagsforskel mellem to kaffesorter. I testen deltog 600 personer. Ved den lejlighed var der 227 af testpersonerne, som udpegede den rigtige kaffekop, altså den, som skiller sig ud.

a) Opstil en nulhypotese og en alternativ hypotese.
b) Redegør for, hvorfor der er tale om et binomialforsøg. Foretag derefter et binomialtest med et signifikansniveau på 5% med henblik på at afgøre, der er smagsforskel på de to kaffesorter eller ej.

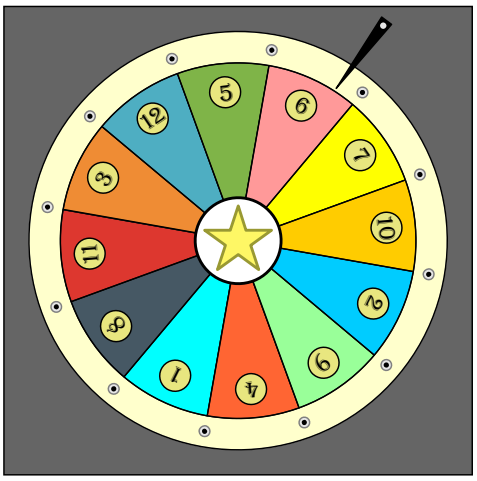
Opgave 407 (Dronningens nytårstale)

Ifølge TNS Gallup så 1676000 danskere dronningens nytårstale den 31. december 2019. Det svarer til en seerprocent på ca. 28,8% i hele landet. En radiovært i Sønderjylland vælger at undersøge, om seerprocenten i Sønderjylland var den samme som på landsplan. Derfor lægger han spørgsmålet op på radiostationens Facebook side, så folk kan respondere, om de overværede talen eller ej. Der blev modtaget 856 svar. Af dem skrev 201 personer, at de havde overværet talen.

a) Redegør hvorfor der er tale om et binomialforsøg. Opstil derefter en nulhypotese og en alternativ hypotese, der skal anvendes til at teste, om seerprocenten til dronningens nytårstale var den samme i Sønderjylland, som i landet som helhed.
b) Benyt et binomialtest med et 5% signifikansniveau til at vurdere, om seerprocenten var den samme i Sønderjylland som i hele landet.
c) Hvad er *stikprøven* og *populationen* i dette binomialtest?
d) Hvilken kritik kan man stille af radioværtens metode til at afklare spørgsmålet?

Opgave 408

På et lykkehjul i et tivoli er der 12 tal. Ifølge indehaveren af boden er alle udfald lige sandsynlige. Topgevinst fås, når pilen rammer 12-tallet.



a) Bestem sandsynligheden for i 70 spil at få netop to 12'ere.
b) Hvad er det mest sandsynlige antal 12'ere i 70 spil på lykkehjulet?
c) Hvad er sandsynligheden for at få højst fire 12'ere i 70 spil?
d) Bestem sandsynligheden for at få flere end otte 12'ere i 70 spil.

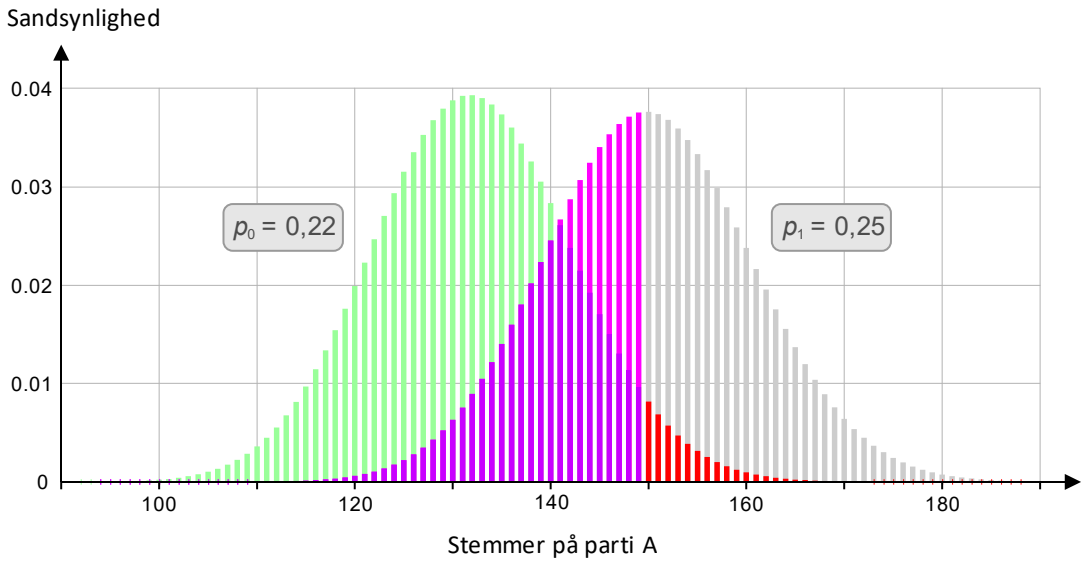
Carl har kigget på spillet i en halv times tid, mens andre har spillet. Han synes, at det virker som om, at 12'eren ikke kommer så ofte ud, som den burde, måske fordi fjederen i pinden op til 12-taller er strammet, tænker han. Til sidst henvender han sig til indehaveren af boden, som dog bedyrer, at der ikke skulle være noget i vejen med lykkehjulet. Han giver endda Carl lov til at teste den efter lukketid. Carl tager imod tilbuddet og gennemfører 750 spil, hvor han får i alt 57 12'ere.

e) Opstil en nulhypotese og en alternativ hypotese, som kan bruges til at teste, om Carls mistanke har hold i virkeligheden eller ej, altså om frekvensen af 12'ere er lavere end den burde være.
f) Benyt et binomialtest med et signifikansniveau på 5% til at undersøge, om frekvensen af 12'ere ved spil på lykkehjulet giver et for lavt tal eller ej.

4.3 Binomialtest: Vigtige grundlæggende erkendelser

Opgave 409 (Type II fejl i binomialtests)

I afsnit 4.3 omtalte vi de såkaldte type I og type II fejl ved binomialtest. Vi kom frem til, at sandsynligheden for, at der forekommer en type I fejl, er lig med signifikansniveauet på α – eller rettere højst lig med α, grundet den diskrete fordeling. Det blev desuden påstået, at sandsynligheden for en type II fejl er mere kompliceret at angive. Det er den, vi vil kigge på i denne opgave. Vi vil tage udgangspunkt i eksempel 4.5 i afsnit 4.2. Men hvad ville en type II fejl da være i den situation? Jo ifølge skemaet side 55 er en type II fejl en, hvor nulhypotesen accepteres, selv om den faktisk er forkert, dvs. at partiet i virkeligheden er gået frem, hvis man så på hele populationen. Det problematiske ved at vurdere sandsynligheden for denne fejl er, at man nu skal antage den alternative hypotese, som ikke bare indeholder én enkelt sandsynlighedsparameter $p_0 = 0{,}22$, men et interval: $p > 0{,}22$. Dermed har vi ikke umiddelbart mulighed for at tegne et pindediagram, for hvilken sandsynlighedsparameter større end 0,22 skalle man vælge? Der er ingen entydig løsning på det. Vi skal i det følgende som et eksempel vælge en mere eller mindre tilfældig værdi større end 0,22 og ud fra den lave et pindediagram: vi vælger $p_1 = 0{,}25$.



Type II fejlen kommer til at forekomme for alle stikprøver, hvor der er under 150 stemmer på partiet A, svarende til, at vi er til venstre for det røde kritiske område for nulhypotesen. Dermed skal vi have bestemt den kumulerede sandsynlighed for binomialfordelingen med antalsparameter n og sandsynlighedsparameter $p_1 = 0{,}25$. Det svarer til at "addere alle de pink/lilla pinde" på figuren ovenfor:

$$bincdf(n, p_1, 149) \;=\; 0{,}4843$$

Altså hele 48,4% sandsynlighed for at begå type II fejl, hvis vi vælger at tage udgangspunkt i sandsynlighedsparameteren $p_1 = 0{,}25$ fra den alternative hypotese! Men nu var valget af sandsynlighedsparameter fra den alternative hypotese som sagt lidt tilfældig.

a) Hvor stor sandsynlighed er der for at begå en type II fejl, hvis udgangspunktet for en sandsynlighedsparameter fra den alternative hypotese er $p_1 = 0{,}27$?

Man vil kunne se, at sandsynligheden for en type II fejl – betegnet β – aftager ret hurtigt, jo højere en værdi for p_1, man vælger. Når β betegner sandsynligheden for at acceptere nulhypotesen, når den er forkert, så må $1-\beta$ svare til sandsynligheden for at afvise nulhypotesen, når den virkelig er forkert (altså en korrekt konklusion!). Man taler om en såkaldt *styrkekurve* (på engelsk: *Powercurve*), som er grafen for $1-\beta$ som funktion af sandsynligheden p_1. Denne kurve karakteriserer, hvor god binomialtesten er til at afvise en nulhypotese, når den virkelig er forkert. Kurven vil blandt andet afhænge af valget af α for signifikansniveauet. Er α valgt lille, bliver β større og omvendt. Der er dog en vigtig størrelse, som også har en indflydelse få β og dermed også $1-\beta$, og det er n! Ikke overraskende vil fejl af type II blive reduceret, når man vælger en større stikprøve.

b) Hvor stor er sandsynligheden for (korrekt) at afvise nulhypotesen, når den er forkert i tilfældet med $p_1 = 0{,}27$ (stadig med $n = 600$)?

c) Undersøg, hvor meget β vil blive reduceret til, hvis vi vælger en stikprøve på 1200 personer i stedet for ovenstående på 600 – med udgangspunkt i $p_1 = 0{,}25$.

4.4 Binomialtest via *p*-værdien

Opgave 410 (Kommunevalg)

Ved forrige kommunevalg i en større by fik partiet P i alt 16,5% af stemmerne. Spørgsmålet er, hvordan det står til med tilslutningen til partiet kort før det kommende kommunevalg. Ved en vælgerundersøgelse blandt 910 tilfældigt valgte indbyggere udtalte 170 vælgere, at de stadig vil stemme på partiet.

a) Opstil en nulhypotese, der kan anvendes til at teste, om tilslutningen til partiet er uændret. Vi anvender et tosidet test, da der ikke er nogen specielle indikationer på hverken fremgang eller tilbagegang i forhold til forrige valg.
b) Hvor mange stemmer ville der forventet være på partiet P, hvis nulhypotesen er korrekt? *Hjælp*: Benyt formlen for middelværdien af en binomialfordelt stokastisk variabel X til at besvare spørgsmålet.
c) Bestem p-værdien for den aktuelle værdi af den stokastiske variabel X på 165. *Hjælp*: Du skal bestemme den kumulerede frekvens $P(X \geq 170)$ og derefter gange med 2, da det er et tosidet test.
d) Hvad er konklusionen: Kan nulhypotesen accepteres eller forkastes på et 5% signifikansniveau?

4.5 Konfidensintervaller for en andel

Opgave 411 (Vælgertilslutning – teoretiske pointer)

I en tilfældig og repræsentativ vælgerundersøgelse blev 650 personer i en jysk by spurgt, om hvilket parti, de agter at stemme på ved det kommende kommunevalg. Et parti K ville i den forbindelse få 181 stemmer.

a) Benyt sætning 4.12 til at give et 95% konfidensinterval for tilslutningen til partiet K.
b) Udtryk sprogligt, hvad dette interval fortæller.

Et andet parti L fik kun 45 stemmer.

c) Bestem 95% konfidensintervallet for tilslutningen til partiet L.
d) Hvad er bredderne af konfidensintervaller for de to partier?

Forestil dig, at man teoretisk set havde udspurgt dobbelt så mange, dvs. 1300 personer, og der ligeledes havde været dobbelt så mange stemmer på partiet K, dvs. 362 stemmer.

e) Hvad ville konfidensintervallet da have været for partiet K? Hvad er bredden?
f) Hvad slutter du om betydningen af det antal personer, der udspørges i vælgerundersøgelser?
g) (Svær) I delspørgsmål d) så vi, at bredden af 95% konfidensintervallet afhænger af den estimerede andel $\hat{p} = k/n$. Vis, at bredden af intervallet er størst, når den estimerede andel er 50%.

Opgave 412 (Rygning)

Ifølge en årsrapport fra 2018 fra Sundhedsstyrelsen var der på landsplan 17% af mændene, som var daglige rygere, mens 7% af mændene var lejlighedsrygere. En statistiker besluttede at undersøge, om situationen var den samme i Sønderjylland hvad angår daglige rygere. På tilfældig vis blev 1045 mænd fra landsdelen udspurgt om deres rygevaner. Ved den lejlighed svarede 195, at de ryger dagligt.

a) Giv på baggrund af stikprøven et estimat for den andel af mændene i Sønderjylland, som er daglige rygere.
b) Udregn et 95% konfidensinterval for andelen af rygende mænd i Sønderjylland.
c) Afgør med udgangspunkt i konfidensintervallet fra b) hvorvidt de Sønderjyske mænd har samme andel af daglige rygere som andelen er på landsplan.

Opgave 413 (Fitness)

I 2007 udgjorde de 16-19-årige medlemmer i kommercielle fitnesscentre 9% af alle medlemmerne i de pågældende centre. Et analyseinstitut satte sig i 2020 for at undersøge, om andelen af 16-19-årige var uændret siden 2007. I den forbindelse kontaktede man på tilfældig vis 1800 medlemmer fra de kommercielle fitnesscentre. Her svarede 209, at de var i alderen 16-19 år.

a) Bestem et estimat for andelen af 16-19-årige medlemmer i de kommercielle fitnesscentre i 2020 – altså bestem $\hat{p}$. Udregn dernæst et 95% konfidensinterval for andelen af 16-19-årige medlemmer i 2020.
b) Afgør med udgangspunkt i konfidensintervallet fra a) hvorvidt andelen af 16-19-årige medlemmer kan siges at være uændret eller ej.

Opgave 414 (Teoretiske pointer)

a) Afbild den statistiske usikkerhed fra sætning 4.12: $\Delta p = 2 \cdot \sqrt{\dfrac{\hat{p} \cdot (1-\hat{p})}{n}}$

som funktion af stikprøvestørrelsen n. Du kan vælge en vilkårlig værdi for $\hat{p}$. Hvad fortæller forløbet af kurven dig om den betydning, som stikprøvens størrelse har?

b) Foretag en fortolkning af grafen fra a). I hvilke situationer kan det bedst betale sig at øge stikprøvestørrelsen: Når stikprøven er lille eller stor?

Opgave 415 (Kvalitetskontrollør)

Søren, som er kvalitetskontrollør, ønsker at danne sig et overblik over, hvor stor en procentdel defekte elementer, som en bestemt maskine producerer. Han udvælger på tilfældig måde 1450 elementer, hvoraf 62 viser sig at være defekte. Bestem et konfidensinterval for andelen af defekte elementer.

Opgave 416 (Kast med mønt)

En mønt kastes 120 gange. Derved fremkom 69 gange krone. Bestem et 95% konfidensinterval for andelen af krone ved kast med mønten. Benyt dette til at vurdere, om der er grund til at mistænke mønten for at være uægte.

Opgave 417 (Matematikeksamen)

Ved en landsdækkende matematikeksamen med flere tusinde studenter udtrak man på tilfældig vis en stikprøve på 185 besvarelser med henblik på at vurdere dumpeprocenten for årets sæt. I blandt de 185 besvarelser var der 41, som dumpede. Bestem et 95% konfidensinterval for andelen af prøver, som dumpede. Efter kritik fra de forrige år, hvor dumpeprocenten var ret høj, ønsker opgavestilleren at dumpeprocenten dette år ikke overstiger 20%. Har opgavestilleren grund til at være nervøs?

Opgave 418 (Markedsanalyse)

For 1 år siden fik en supermarkedskæde gennemført en markedsanalyse, som fortalte, at 28% af befolkningen i det omkringliggende område regelmæssigt handlede ind hos dem. For at se, hvordan det står til i år, har man igen spurgt en tilfældig gruppe af personer i området. Den nye stikprøve giver en estimeret andel af regelmæssige kunder på 32%. Konfidensintervallet er så smalt, at man konkluderer, at andelen af regelmæssige kunder *ikke* er uændret. Hvor stor må stikprøvestørrelsen mindst have været? *Hjælp*: Husk at jo større n, jo mindre statistisk usikkerhed. Kig på udtrykket for den statistiske usikkerhed.

4.6 Konfidensinterval for et gennemsnit

Opgave 419 (Konfidensinterval for middelværdien - projekt)

I afsnit 4.6 kiggede vi på et eksempel med længder af en bestemt type fisk. Vi ønskede at bestemme gennemsnitslængden af denne type fisk i søen samt et konfidensinterval for gennemsnitslængden ved hjælp af data hentet fra en stikprøve. I det følgende skal du benytte sætning 4.18 til igennem nogle trin at nå frem til at give et konfidensinterval for gennemsnitslængden af en bestemt type fisk i en sø. Man har på tilfældig måde udtaget en stikprøve på 50 fisk af en bestemt art fra flere steder i søen og målt fiskenes længder. Det gav følgende længder, regnet i cm:

53, 76, 70, 68, 52, 67, 68, 58, 80, 61, 66, 63, 73, 65, 57, 71, 51, 85, 52, 64, 63, 61, 60, 70, 71, 66, 67, 49, 51, 75, 77, 69, 85, 69, 59, 59, 80, 71, 81, 86, 53, 55, 65, 62, 64, 54, 64, 73, 63, 75.

a) Bestem på grundlag af stikprøven den empiriske middelværdi for længden af fiskene.
b) Bestem stikprøvespredningen s.
c) Bestem standardfejlen på middelværdien.
d) Udregn konfidensintervallet for fiskenes gennemsnitslængde og formuler omhyggeligt, hvad dette interval fortæller noget om.
e) Hvad ville der ske med størrelserne $\overline{x}, s, s_{\overline{x}}$ samt konfidensintervallet, hvis stikprøven var 4 gange så stor? Tænk for eksempel på hvis hver af de oprindelige længdemålinger havde 3 dubletter. Hvilken betydning konkluderer du, at stikprøvestørrelsen har på usikkerheden i bestemmelsen af fiskene gennemsnitslængde?

Det gode ved situationen i sætning 4.18 er, at den gælder for stikprøver, hvor elementerne ikke behøver være værdier af stokastiske $X_1, X_2, \ldots, X_n$ med en bestemt fordeling, bare de har den samme fordeling. Ulempen er, at sætning 4.18 kun gælder approksimativt,

ligesom stikprøven skal have en ikke helt lille størrelse. Men hvad gør man så, hvis man har en lille stikprøve? Ja så har man en anden variant af sætningen, men kun hvis man til gengæld ved at X_i'erne er normalfordelte med samme middelværdi μ og spredning σ. Ligesom soldaters højde er normalfordelte, er det også rimeligt at antage, at fisks længde er det. Den nye sætning involverer så også den såkaldte *Student t*-fordeling, som blev opdaget af den britiske statistiker *William Sealy Gosset* (1876-1937). Han opdagede, at mens $(\overline{X}-\mu)/(\sigma/\sqrt{n})$ er eksakt normalfordelt, når X_i'erne er det, så er det ikke tilfældet, når man udskifter den ukendte spredning σ med estimatoren *S* for stikprøvespredningen. I det tilfælde vil $(\overline{X}-\mu)/(S/\sqrt{n})$, have en helt ny fordeling. Den fik navnet *Student-t*-fordelingen eller bare *t*-fordelingen. Det skal nævnes, at når $n\to\infty$ vil *t*-fordelingen konvergere mod en normalfordeling. Vi angiver uden bevis:

Sætning (Konfidensinterval for middelværdien – ved normalfordelte elementer med ukendt varians)

Givet en simpel tilfældig stikprøve bestående af *n* elementer: $\{x_1, x_2, \ldots, x_n\}$, som er værdier af de stokastiske variable $X_1, X_2, \ldots, X_n$, alle med samme normalfordeling med middelværdi μ og ukendt spredning σ. Da er det eksakte 95%-konfidensinterval-givet ved følgende formel:

$$\left[\overline{x}-t_{0.975}(n-1)\cdot\frac{s}{\sqrt{n}},\ \overline{x}+t_{0.975}(n-1)\cdot\frac{s}{\sqrt{n}}\right]$$

hvor *s* er stikprøvespredningen fra sætning 4.18 og hvor $t_{0.975}(n-1)$ er 0,975-fraktilen for *t*-fordelingen med $n-1$ frihedsgrader.

I det følgende skal du give et estimat for gennemsnitslængden μ af en bestemt type fisk i en sø og bestemme det tilhørende 95%-konfidensinterval, under antagelse af, at fiskenes længde er normalfordelte med middelværdi μ og ukendt spredning σ. Du skal benytte sætningen ovenfor. Undersøg selv, hvordan man i dit regneværktøj bestemmer de involverede fraktiler for *t*-fordelingen. Stikprøven indeholdt følgende fiskelængder i cm:

83, 75, 61, 84, 66, 77, 64, 62, 52, 74, 73, 66, 63, 74, 58, 62, 65, 69.

f) Bestem stikprøvegennemsnittet $\overline{x}$ og stikprøvespredningen *s*.
g) Bestem værdien $t_{0.975}(17)$ med $n-1=18-1=17$ frihedsgrader i dit CAS-værktøj.
h) Bestem konfidensintervallet for fiskenes gennemsnitslængde og formuler omhyggeligt, hvad dette interval fortæller noget om.

5.2 Lineær regression og mindste kvadraters metode

Opgave 501

I bemærkning 5.3 blev det påstået, at hældningskoefficienten a for regressionslinjen i sætning 5.1 kan skrives på en alternativ måde:

$$a = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = \frac{\overline{x\cdot y}-\overline{x}\cdot\overline{y}}{\overline{x^2}-\overline{x}^2}$$

hvori der i udtrykket til højre indgår flere middelværdier:

$$\overline{x\cdot y} = \sum_{i=1}^{n} x_i\cdot y_i\,,\quad \overline{x} = \sum_{i=1}^{n} x_i\,,\quad \overline{y} = \sum_{i=1}^{n} y_i\,,\quad \overline{x^2} = \sum_{i=1}^{n} x_i^2$$

a) (Noget teknisk svær). Vis identiteten. *Hjælp*: Gang parenteserne i udtrykket på venstre side ud og sæt alt det udenfor summen, som ikke afhænger af i. Reducer så meget som muligt. Indsæt middelværdierne i udtrykket på højre side. Vis at venstre og højre side reducerer til det samme.

b) I tabellen nedenfor er givet en række punkter. Bestem regressionslinjen $y = a\cdot x + b$ ved at benytte den alternative formel for a ovenfor, samt formlen for b fra sætning 5.1, dvs. $b = \overline{y} - a\cdot\overline{x}$.

x	-5	-3	-2	1	4	5	7	10	12	15
y	-1,0	0,3	3,0	6,0	4,6	9,0	5,8	10,2	12,3	12,7

Kontroller til slut, om du har regnet rigtigt, ved at bruge dit CAS-værktøj til automatisk at udføre lineær regression.

Opgave 502 (Residualer)

I tabellen nedenfor er angivet 5 datapunkter. Når man udfører lineær regression på disse punkter, viser det sig, at regressionslinjen får følgende forskrift: $f(x) = 1{,}7x + 3{,}3$.

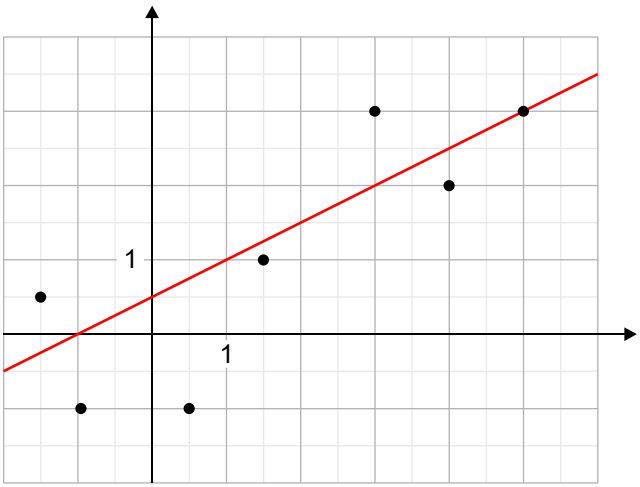
x_i	-1	0	1	2	3
y_i	2	3	5	6	9
$f(x_i)$					
r_i					

a) Bestem funktionsværdierne og residualerne, dvs. bestem værdierne i de tomme rubrikker i tabellen ovenfor.

b) Vis at summen af residualerne giver 0, som sætning 5.9 siger altid er tilfældet.

Opgave 503 (Residualer)

Figuren herunder viser en del af en række datapunkter med tilhørende regressionslinje.



Aflæs de forskellige værdier og udfyld skemaet nedenfor.

x_i							
y_i							
$f(x_i)$							
r_i							

Opgave 504 (Residualer)

Der er foretaget lineær regression på 10 datapunkter (x_i, y_i), hvilket gav følgende forskrift for den lineære regressionsfunktion: $f(x) = 0{,}7x - 2{,}1$. I nedenstående tabel er der udvalgt fire af x-værdierne fra datapunkterne.

x_i	-2	3	6	10
y_i	–5,3		2,9	
$f(x_i)$				
r_i		1,6		0,8

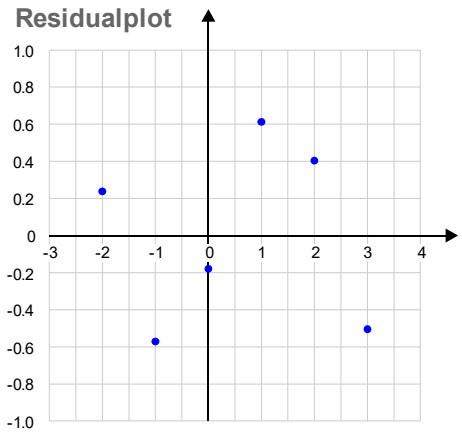
a) Udregn funktionsværdierne i de fire x-værdier og indsæt dem i tabellen.
b) Udfyld resten af de tomme felter i tabellen.

Opgave 505 (Residualer)

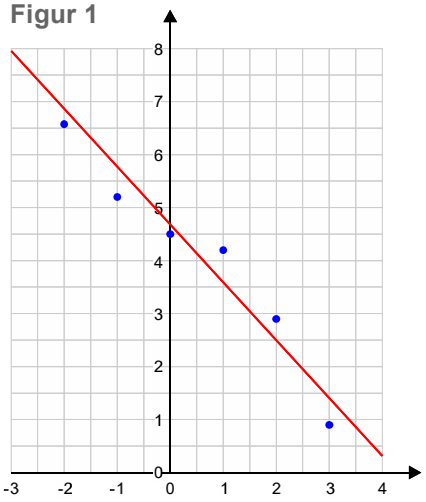
Der er foretaget lineær regression på en række datapunkter. Den lineære regressionsfunktion betegnes $f(x)$. I datapunktet $(3,5)$ er residualet –3. Bestem $f(3)$.

Opgave 506 (Residualer)

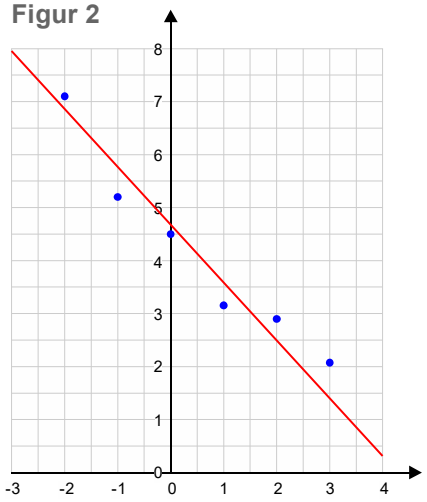
Grafen øverst er residualplottet for ét af de fire regressionsplot nedenfor. Hvilket et er det? Husk at argumentere, gerne ved at udelukke.



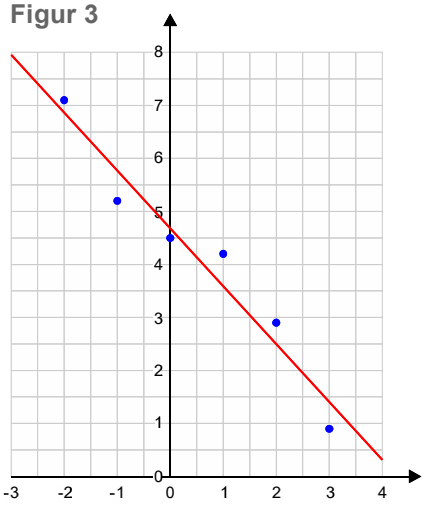
Figur 1



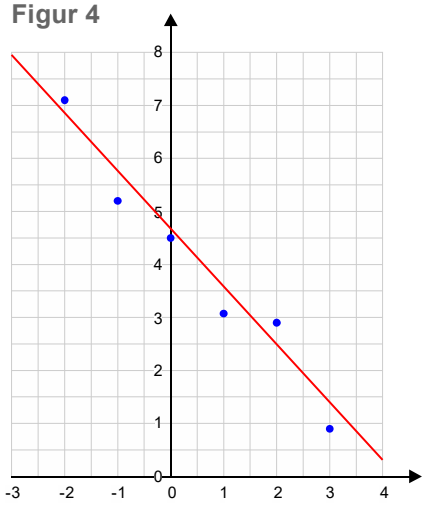
Figur 2



Figur 3

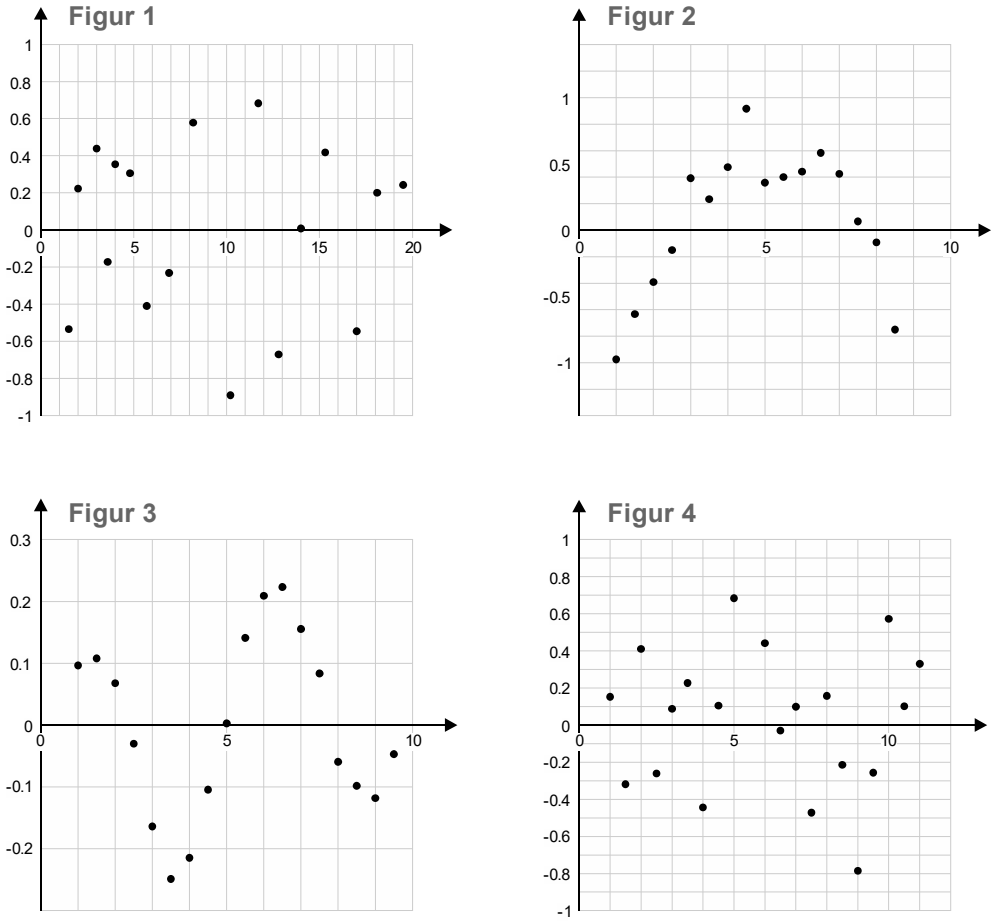


Figur 4



Opgave 507 (Residualer)

Betragt de fire residualplot nedenfor. De stammer alle fra lineære regressioner. Vurder for hvert af tilfældene, om der er tale om en lineær sammenhæng, eller om det tyder på en systematisk afvigelse fra en lineær sammenhæng.



5.3-5.4 Den simple lineære regressionsmodel m.m.

Opgave 508 (Opvarmning af vand)

I en gymnasieklasse udføres forsøg med en elkedel. Man vil undersøge, om temperaturen af 1,2 liter vand vokser lineært som funktion af tiden. På næste side kan findes de 20 sammenhørende værdier af tiden i sekunder og temperaturen i °C.

a) Foretag lineær regression på de 20 datapunkter.
b) Lav et residualplot med dit CAS-værktøj og redegør for, at der er tale om en systematisk afvigelse fra en lineær sammenhæng.
c) Udregn residualerne. Hvor stort er det største og mindste residual, numerisk set?

t (sek)	20	30	40	50	60	70	80	90	100	110
T (°C)	24,8	28,7	32,7	36,5	40,5	44,2	48,0	51,9	55,7	59,4

t (sek)	120	130	140	150	160	170	180	190	200	210
T (°C)	63,4	67,2	71,0	74,8	78,1	80,7	83,4	86,8	88,6	89,4

Årsagen til, at vi må afvise en lineær sammenhæng, er især, at der strømmer meget energi ud af elkedlen, når vandet er blevet meget varmt. Man ser også damp. Konsekvensen er, at vandets temperatur ikke vokser så hurtigt, når temperaturen nærmer sig kogepunktet. Vi vil dog undersøge, om der er en lineær sammenhæng mellem tid og temperatur i de første 140 sekunder, altså indtil vandets temperatur har nået 71,0°C.

d) Foretag lineær regression på de første 13 datapunkter og angiv a og b.
e) Lav et residualplot for de første 13 datapunkter.

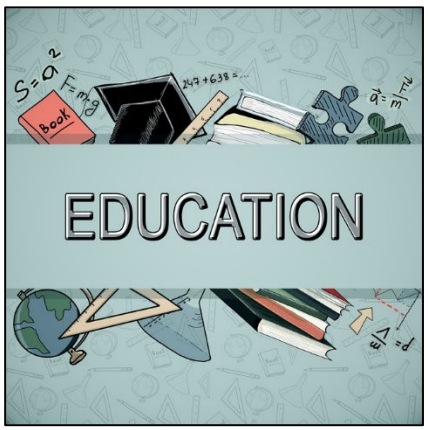
Nu skulle det se bedre ud. Residualplottet er ikke perfekt, men værdierne af residualerne er nu så små, så vi med rimelighed kan godtage en lineær sammenhæng. Det viser en inspektion af grafen for den lineære regression fra d) også.

f) Benyt regressionslinjen fra d) til at give et bud på, hvad vandets temperatur vil være efter 36 sekunder.
g) Anvend modellen til at forudsige, hvornår vandets temperatur har nået 50°C.

Opgave 509 (Lineær regressionsmodel: Uddannelse - Indkomst)

Vi ser i denne opgave på lønmodtagere og deres indkomst. Det er velkendt, at jo længere uddannelse en person har, jo mere vil denne få i løn – gennemsnitligt. I det følgende ser vi på fiktive data: I et land har man taget en stikprøve i befolkningen for at undersøge sammenhængen mellem det antal år, som personen har været i uddannelse og den månedsløn, som lønmodtageren opnår. Data findes i filen:

uddannelse-indkomst.xlsx.



Uddannelse (år)	11	19	...	13	18	10
Indkomst (kr.)	20509	38567		23678	35463	16506

Dele af data fra filen uddannelse-indkomst.xlsx

I en model antages det, at sammenhængen mellem længden af en persons uddannelse og dennes månedsløn før skat er beskrevet ved en lineær funktion

$$f(x) = a \cdot x + b$$

hvor x er den samlede uddannelsestid (målt i år), og $f(x)$ betegner månedslønnen (målt i kr.).

a) Bestem tallene a og b.
b) Bestem residualspredningen s.
c) Lav et residualplot.
d) Bestem, hvor mange procent af residualerne, der ligger indenfor intervallet $[-2s, 2s]$.

Opgave 510 (Lineær regressionsmodel: Højde-FEV1)

Ved et idrætsarrangement foretager deltagerne, som alle er unge mænd i midten af 20'erne, en lungetest. Der er tale om en såkaldt FEV1-test, hvor personen først trækker vejret dybt ind og derefter puster luften ud igennem et mundstykke tilkoblet et apparat. Apparatet måler, hvor meget luft personen puster ud i løbet af 1 sekund (målt i liter). FEV1 er *Forced Expiratory Volume*. På dansk: *Forceret udåndingsvolumen*. Lungetesten har til formål at undersøge sammenhængen mellem en persons højde og dennes FEV1.

Højde (cm)	191	182	...	178	173	173
FEV1 (L)	5,09	4,46		4,02	4,14	4,81

Dele af data fra filen højde-FEV1.xlsx

I en model antages det, at sammenhængen mellem personens højde og FEV1 værdi er beskrevet ved en lineær funktion

$$f(x) = a \cdot x + b$$

hvor x er personens højde (målt i cm), og $f(x)$ betegner personens FEV1-værdi (målt i liter).

a) Foretag lineær regression og bestem a og b i modellen.
b) Tegn et residualplot og bestem residualspredningen.
c) Benyt residualplottet og residualspredningen til at vurdere modellens anvendelighed til at beskrive udviklingen.
d) Gør rede for at mindst 95% af residualerne ligger i intervallet $[-2s, 2s]$.

5.5 Ekstra: Konfidensinterval for hældning

Opgave 511 (Fortsættelse af opgave 509 med Uddannelse-Indkomst)

Læseren antages allerede at have løst opgave 509. Her er en række tillægsspørgsmål:

a) Gør rede for, at residualerne med god tilnærmelse kan siges at være normalfordelte.
b) Angiv middelværdi og spredning for de normalfordelte residualer.
c) Bestem et 95% konfidensinterval for hældningen a.

Opgave 512 (Fortsættelse af opgave 510 med Højde-FEV1)

Læseren antages allerede at have løst opgave 510. Her er en række tillægsspørgsmål:

a) Bestem et 95% konfidensinterval for hældningen a, og benyt dette til at afgøre, om sammenhængen mellem en persons højde og FEV1 er voksende.

Opgave 513 (Hvilken type sammenhæng?)

Lad os sige, at vi har udtrukket en stikprøve $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ af data fra en simpel lineær regressionsmodel og efterfølgende udregnet regressionslinjen $y = a \cdot x + b$ og et 95% konfidensinterval for hældningskoefficienten a. Afgør i hvert af tilfældene nedenfor, om det er rimeligt at antage, at der er tale om en *voksende lineær sammenhæng*, en *aftagende lineær sammenhæng* eller *ingen sammenhæng*.

a) $[76{,}2; 85{,}7]$ b) $[-0{,}412; 3{,}75]$ c) $[-230{,}1; -187{,}6]$ d) $[8{,}2; 8{,}7]$

Opgave 514 (Er der en lineær sammenhæng eller ej?)

Excel filen *lineær_sammenhæng_eller_ej.xlsx* indeholder 250 datapunkter.

x	37,9	52,3	...	43,4	48,8	45,0
y	31,5	22,2		26,9	18,8	20,0

Dele af data fra filen lineær_sammenhæng_eller_ej.xlsx

a) Importer datapunkterne i dit CAS-værktøj og foretag lineær regression. Angiv a og b i forskriften for den lineære funktion.
b) Lav et residualplot og vis, at residualerne er normalfordelte. Angiv desuden en værdi for residualspredningen.
c) Bestem et 95% konfidensinterval for hældningen og afgør, om der mon er tale om en lineær sammenhæng eller ej.

Opgave 515 (Markedsanalyse for supermarked)

En markedsanalytiker, som er hyret af et større supermarked, ønsker at analysere sammenhængen mellem prisen og efterspørgslen (salget) for en bestemt fødevare. Som en model antager man, at sammenhængen kan beskrives ved en lineær funktion:

$$f(x) = a \cdot x + b$$

hvor x betegner stykprisen på varen i kr. og $f(x)$ er antallet af solgte styk af varen i løbet af et døgn.

Undersøgelsen foretages ved at variere prisen på den pågældende vare, som desuden er gjort ekstra synlig ved at have en fremtrædende plads. For hver ny uge ændres varens pris til en anden. For hver af de 7 dage i ugen gøres det op, hvor mange enheder, der er solgt. Tabellen herunder viser de målte data.

Pris (kr.)	17,50	17,50	...	26,50	26,50	26,50
Antal	152	154		89	100	95

Dele af data fra filen pris_salg.xlsx

a) Foretag lineær regression på data og angiv værdier for a og b.
b) Gør rede for, at residualerne med god tilnærmelse kan siges at være normalfordelte.
c) Lav et residualplot og bestem residualspredningen s.
d) Hvor stor en procentdel af residualerne ligger i intervallet $[-2s, 2s]$?
e) Angiv værdien for hældningskoefficienten a og giv en fortolkning af den.
f) Angiv et 95% konfidensinterval for hældningskoefficienten a.
g) Hvor mange styk af varen vil supermarkedet i gennemsnit sælge pr. dag, hvis stykprisen sættes til 22 kr.?

Opgave 516 (Projekt med Galton data: Forældres højde og deres børns højde)

I 1880'erne studerede den britiske statistiker Sir Francis Galton (1822-1911) sammenhængen mellem forældres højde og deres voksne børns højde. Han var klar over, at højderne af mænd og kvinder hver for sig er normalfordelte, men hvor normalfordelingerne har forskellig middelværdi. Begge forældres højde kan tænkes at have en indvirkning på afkommets højde, og afkommets højde må afhænge af kønnet. For at tage højde for begge forældes højde og kvindens lavere gennemsnitshøjde, indførte Galton en "mid-parental height". Vi vil kalde den *midlet forældrehøjde*. Den er defineret ved, at man ganger moderens højde med 1,08 og derefter tager gennemsnittet mellem denne justerede kvindehøjde og så faderens højde:

$$h_{\text{midlet forældrehøjde}} = \tfrac{1}{2} \cdot \left(h_{\text{fader}} + 1{,}08 \cdot h_{\text{moder}} \right)$$

På tilsvarende måde indførte han en *justeret børnehøjde* derved, at hvis det voksne barn er en kvinde, så ganges dets højde med 1,08:

$$h_{\text{justeret barnehøjde}} = \begin{cases} h_{\text{barn}} & \text{hvis barnet er en mand} \\ 1{,}08 \cdot h_{\text{barn}} & \text{hvis barnet er en kvinde} \end{cases}$$

I det følgende skal du bruge Galtons originale data, som befinde sig i Excel filen:

Galton_højder.xlsx

Alle højder er i tommer (inches). 1 inch = 2,54 cm. De første fem kolonner i filen er Galtons originale. Derudover er der i søjlerne G og H tilføjet henholdsvis de midlede forældrehøjder og de justerede børnehøjder, udregnet efter formlerne ovenfor.

a) Importer i dit CAS-værktøj søjlerne G og H fra Excel filen og giv dem navnene henholdsvis X og Y. Vis, ved at lave et QQ-plot, at de hver især indeholder data, som med god tilnærmelse kan siges at være normalfordelte.

b) Foretag lineær regression på de sammenhørende værdier af de midlede forældehøjder og de tilhørende justerede børnehøjder. Noter hældningskoefficienten a for regressionslinjen ned sammen med forklaringsgraden R^2.

c) Lav et QQ-plot af residualerne i den lineære regression for at se, om de er normalfordelte. Tyder det på, at vi har at gøre med en simpel lineær regressionsmodel?

d) Et forældrepar, hvor mandens højde er 75,5 tommer og kvindens højde er 66.2 tommer, får sammen en dreng. Udregn den midlede forældrehøjde. Udnyt den, sammen med regressionslinjen fra b), til at give et bud på den højde, som en dreng fra disse forældre i gennemsnit vil få.

e) Et andet forældrepar, hvor mandens og kvindens højde er henholdsvis 68,9 tommer og 64,2 tommer, får sammen en pige. Udregn først den midlede forældrehøjde. Benyt derefter regressionslinjen fra b), til at give et bud på den højde, som en pige fra disse forældre i gennemsnit vil få. NB! Husk at justere barnets højde tilbage med en faktor 1,08, da der er tale om en pige!

Vi har i a) set, at de stokastiske variable X og Y, som repræsenterer henholdsvis de midlede forældrehøjder og de justerede børnehøjder, hver især er normalfordelte (marginale fordelinger). Tilsammen udgør de en todimensional normalfordeling. Af den lineære regression i b) opdagede vi, at der er en sammenhæng mellem den midlede forældrehøjde og den justerede barnehøjde. Højere forældre får gennemgående også højere børn. Ikke altid, men i gennemsnit! Spørgsmålet er *med hvor stor effekt* eller *styrke* forældrenes højde slår igennem hos børnene. Man kunne måske tænke sig at bruge hældningskoefficienten fra den lineære regression til at afgøre det. Det virker nogenlunde, når man som her har de samme størrelser og de samme enheder på akserne. Generelt set er hældningskoefficienten dog afhængig af de anvendte enheder. Der er en mere sigende størrelse: *Korrelationskoefficenten*. Rent formelt er den defineret ud fra de stokastiske variable X og Y for henholdsvis de midlede forældrehøjder og de justerede børnehøjder via den såkaldte *kovarians*. Det vil føre for vidt at komme ind på den her. Vi holder tingene nede på et

mere intuitivt plan. Vi vil estimere korrelationskoefficienten ρ via stikprøven bestående af sammenhørende værdier af X og Y fra de to søjler i Excel filen. Den *empiriske korrelationskoefficient*, som er *uafhængig* af enheder, er defineret ved:

Definition

Den *empiriske korrelationskoefficient*, som beskriver samvariationen mellem x- og y-værdierne i sættet af datapunkter: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, er defineret ved:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \overline{x})\cdot(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{1}$$

hvor $\overline{x}$ er gennemsnittet af alle x-værdierne og $\overline{y}$ er gennemsnittet af alle y-værdierne i datasættet.

Man kan vise følgende alternative formel for ρ (se eventuelt [L2]):

$$\rho = a\cdot\frac{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{2}$$

hvor a er hældningskoefficienten for regressionslinjen. Begrebet *stikprøvespredning* eller *empirisk spredning* er defineret i Appendiks A. Stikprøvespredningerne s_x og s_y for henholdsvis x-værdierne og y-værdierne er defineret ved:

$$s_x = \sqrt{\frac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i - \overline{x})^2} \quad \text{og} \quad s_y = \sqrt{\frac{1}{n-1}\cdot\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{3}$$

De betegnes også ofte *standardafvigelser* i henholdsvis x og y. Det giver anledning til følgende sætning:

Sætning

Givet et sæt af datapunkter $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Sammenhængen mellem hældningskoefficienten a for regressionslinjen hørende til datasættet og den empiriske korrelationskoefficient ρ, er følgende:

$$\rho = \frac{s_x}{s_y}\cdot a \quad \Leftrightarrow \quad a = \rho\cdot\frac{s_y}{s_x} \tag{4}$$

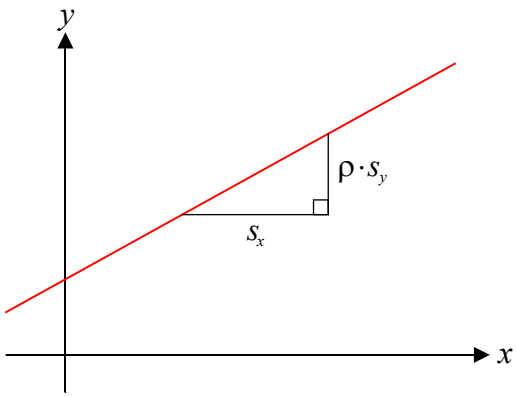
Bevis: Fås umiddelbart ud fra (2) og (3), idet de to konstanter $1/(n-1)$ i (3) forsvinder, når der divideres i (2). □

f) Benyt den alternative formel (2) eller sammenhængen i (4) ovenfor til at vise, at estimatet for korrelationskoefficienten er ca. 0,51, dvs. $\rho \approx 0{,}51$. Giv gerne flere cifre.

Man kan vise, at der er følgende sammenhæng mellem den empiriske korrelationskoefficient og forklaringsgraden R^2 for regressionslinjen: $\rho = \pm\sqrt{R^2}$. Her skal plus bruges, hvis hældningskoefficienten for regressionslinjen er positiv, mens minus skal bruges, hvis hældningskoefficienten er negativ.

g) Vis, at ovenstående påståede sammenhæng mellem ρ og R^2 stemmer i denne opgave, idet du bruger værdierne udregnet i f) og b).

Sætningen vil ikke blive bevist her. Den interesserede læser henvises til min note [L2].



Hældningskoefficienten *a* i den simple lineære regressionsmodel angiver jo, hvor meget *y* vokser med *i middel*, når *x* vokser med 1. Men den er som nævnt enhedsafhængig. Det er bedre at regne i enheder af standardafvigelser! Her siger sætningen ovenfor, at hvis *x* vokser med 1 standardafvigelse i *x*, så vokser *y* i middel med ρ standardafvigelser i *y*. Ifølge f) er korrelationskoefficienten i vores situation lig med ca. 0.51. Det betyder, at hvis den midlede forældrehøjde er 1 standardafvigelse s_x *over* gennemsnitshøjden i befolkningen, så vil forældrenes barn i middel få en justeret højde, som er 0,51 standardafvigelser s_y *over* gennemsnitshøjden. Men husk at alt er estimater ud fra stikprøven! Er forældrenes midlede forældrehøjde derimod for eksempel 0,5 standardafvigelser *under* befolkningens gennemsnitshøjde, så vil børnenes justerede højde dermed være $0{,}5 \cdot 0{,}51 = 0{,}255$ standardafvigelser s_y *under* befolkningens gennemsnitshøjde, som er $\overline{y}$. Korrelationskoefficienten fortæller altså, hvor stor effekt eller styrke forældrenes højde har på afkommets højde, målt i enheder af standardafvigelser, hvilket generelt set er mere sigende. Når vi som her har to størrelser, som er normalfordelte, så ved vi at antallet af standardafvigelser fra middelværdien har en ganske bestemt oversættelse til procenter.

Man kan vise, at korrelationskoefficienten altid vil give et tal mellem –1 og 1. Det betyder så også, at børnenes justerede højde i middel *højst* vil være lige så mange standardafvigelser over/under gennemsnitshøjden, som forældrenes midlede forældrehøjde er henholdsvis over/under gennemsnitshøjden. Der er altså en tendens til, at børns højde går mere *ind mod middelværdien*. Høje forældre vil i gennemsnit få høje børn, men ikke så høje som dem selv. Forældre, hvis højde er under gennemsnitshøjden, vil i gennemsnit få børn med

en højde under gennemsnithøjden, men ikke så langt under gennemsnitshøjden som dem selv. Alt dette er vel at mærke i *gennemsnit*! Galton observerede dette fænomen og kaldte det for *regression towards the mean*. Heraf også betegnelsen *regression*, som kommer af et ord på latin, som betyder "gå tilbage".

h) Bestem værdierne for $\overline{x}$, $\overline{y}$, s_x og s_y i den konkrete situation.

i) Et forældrepar har en midlet forældrehøjde, som er 1,5 standardafvigelser s_x *over* gennemsnitshøjden. Parret føder en dreng. Hvor mange standardafvigelser s_y vil drengens voksenhøjde (i middel) være over gennemsnitshøjden $\overline{y}$?

Bemærkning 1: Se evt. mit dokument [L2] for flere tekniske detaljer.

Bemærkning 2: Det skal desuden tilføjes, at faktoren 1,08 afhænger af populationen og nok også af tidspunktet i historien. Galton estimerede faktoren ud fra stikprøven.

Bemærkning 3: Læseren undres måske over, at værdierne for s_x og s_y fra spørgsmål h) ikke er næsten lige store. Der er jo tale om menneskehøjders variation i begge tilfælde. Forklaringen er, at x-værdierne jo er midlede forældrehøjder. Når man midler, aftager spredningen.

Bemærkning 4: Efter Galton i 1880'erne havde gennemført sine undersøgelser med afkommets højde, studerede en anden stor britisk statistiker, nemlig *Karl Pearson* (1857-1936) i begyndelsen af 1900-tallet sammenhængen mellem fædres højde og deres sønners højde. Du kan evt. forsøge at finde data fra det på Internettet.

Litteratur

[1] Richard J. Larsen, Morris L. Marx. *An Introduction to Mathematical Statistics and Its Applications*. 5th Edition, Prentice Hall, 2012.

[2] Joseph K. Blitzstein, Jessica Hwang. *Introduction to Probability*. CRC Press, 2015.

[3] Dimitri P. Bertsekas, John N. Tsitsiklis. *Introduction to Probability*. Athena Scientific, 2008.

[4] David Freedman, Robert Pisani, Roger Purves. *Statistics*. Fourth Edition. W. W. Norton & Company, 2007.

[5] Nikolaj Malchow-Møller og Allan Würtz. *Indblik i statistik – for samfundsvidenskaberne*. 2. udgave. Hans Reitzels Forlag, 2014.

[6] Susanne Ditlevsen, Helle Sørensen. *Introduktion til Statistik*. 4. udgave. Institut for Matematiske Fag, Københavns Universitet, 2015.

[7] Katja Kofoed Svan, Olav Lyndrup. *Sandsynlighedsregning og statistik med binomialfordelingen*. Officielt undervisningsmateriale i matematik fra Undervisningsministeriet, januar 2019.

[8] Axel Bertelsen. *Statistik med matematik*. Systime, 2005.

[9] Alan Agresti, Barbara Finlay. *Statistical Methods for the Social Sciences*. Third Edition, Prentice Hall, 1997.

Hjemmesider

[L1] https://www.matematikfysik.dk/mat/noter_tillaeg/tillaeg_stikproevevariansen.pdf
(Hvorfor $n-1$ i stikprøvevariansen?)

[L2] https://www.matematikfysik.dk/mat/noter_tillaeg/tillaeg_lineaer_regression_kvadratsummer_forklaringsgrad.pdf
(Lineær regression, kvadratsummer og forklaringsgrad)

Billedliste

Side 15: ©iStock.com/GeorgisArt (Blaise Pascal)
Side 15: ©iStock.com/traveler1116 (Simon-Pierre Laplace)
Side 29: ©iStock.com/IUshakovsky (Casino)
Side 35: ©iStock.com/jaroon (Baby)
Side 42: ©Thinkstock/Drazen Zigic (Kvalitetskontrol af komponenter)
Side 44: ©iStock.com/Elenathewise (Mand ved maskine)
Side 51: ©iStock.com/3pod (Valg)
Side 63: ©iStock.com/ Drazen Zigic (Pakkebude)
Side 78: ©iStock.com/a_namenko (Frugt og grønt)
Side 89: ©iStock.com/jorgeantonio (Frugtplantage)
Side 90: ©iStock.com/Hyrma (Gulerødder)
Side 113: ©iStock.com/Wpadington (Betting)
Side 117: ©iStock.com/kadmy (bilkomponent)
Side 118: ©iStock.com/Ankorlight (Tulipaner)
Side 119: ©iStock.com/JFsPic (Lotteri)
Side 120: ©iStock.com/bluebeat76 (Kattemad)
Side 122: Pudelek [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)] (Den kongelige livgarde på Amalienborg)
Side 128: By Octavius [Public domain], via Wikimedia Commons (Sir Francis Galton)
Side 128: ©iStock.com/GeorgiosArt (Blaise Pascal)
Side 133: Erikeltic at English Wikipedia / CC BY-SA (https://creativecommons.org/licenses/by-sa/3.0) (Tre Labrador Retriever hunde)
Side 138: ©iStock.com/dusanpetkovic (Fitness)
Side 139: ©iStock.com/a_namenko (Frugt og grønt)
Side 140: ©iStock.com/suzanna (Fisk)
Side 146: ©iStock.com/peshkov (Education)