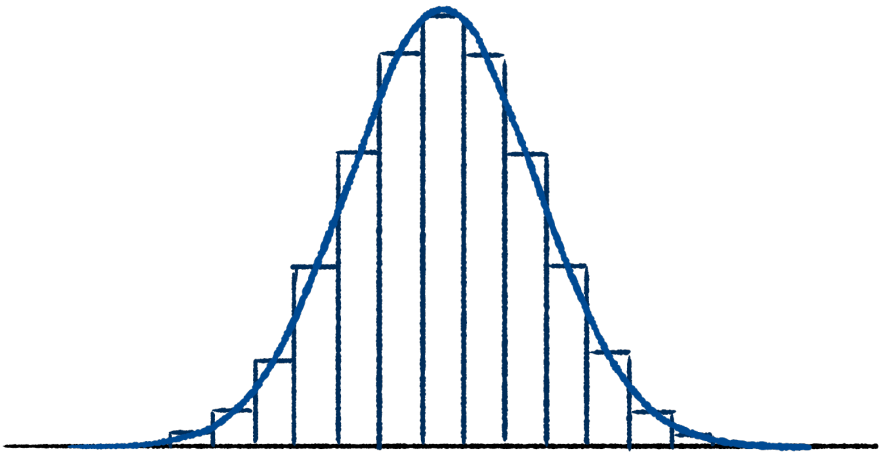
Sandsynlighedsregning

Version 0.95
3. august 2020



Mike Vandal Auerbach
www.mathematicus.dk

Sandsynlighedsregning
Version 0.95, 2020

Disse noter dækker kernestoffet i sandsynlighedsregning på stx A- og B-niveau efter gymnasiereformen 2017.

Enkelte emner fra statistikken er taget med (se afsnittene om binomialtest og stikprøver) da disse bedst kan forklares ud fra sandsynlighedsteoretiske overvejelser.

Disse noter er skrevet til matematikundervisning på stx og må frit anvendes til ikke-kommercielle formål.

Noterne er skrevet vha. tekstformateringsprogrammet LaTeX, se www.tug.org og www.miktex.org. Figurer og diagrammer er fremstillet i *pgf/TikZ*, se www.ctan.org/pkg/pgf.

Disse og andre noter kan downloades fra www.mathematicus.dk.

Mike Vandal Auerbach, 2020

Indhold

Kombinatorik 1

Sandsynlighedsregning er en gren af matematikken, der beskæftiger sig med den kunst at sætte målbare tal på tilfældige fænomener.

Det kunne f.eks. være

- Resultatet af et kast med en terning.
- Gevinsten på et skrabelod.
- Højden af en tilfældigt valgt person.

Sandsynlighederne er en beskrivelse af, hvor ofte et bestemt udfald af eksperimentet forekommer. F.eks. hvor ofte man får 5'ere i et kast med en terning.

1.1 På hvor mange måder ...?

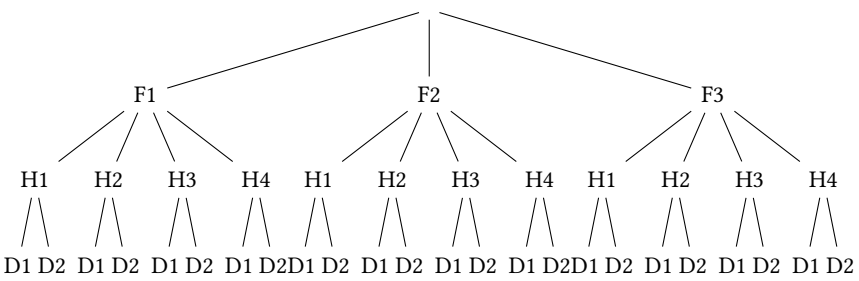
Hvis man skal regne på sandsynligheder, kan det være nyttigt at have formler der gør, at man kan besvare spørgsmålet »på hvor mange måder kan ... lade sig gøre?« Man kan altid tælle sig frem ved at skrive alle mulighederne op, men i praksis kan det blive meget besværligt – især hvis der er mange valgmuligheder.

Man kan starte med følgende konkrete eksempel:

Eksempel 1.1 På en bestemt restaurant kan man vælge mellem 3 forretter, 4 hovedretter og 2 desserter. Hvis en menu består af en forret, en hovedret og en dessert, hvor mange forskellige menuer kan man så sammensætte?

Tegner man et tælletræ over dette vil det se ud som på figur 1.1.

Ved at tælle de nederste grene på træet finder man frem til, at der er 24 forskellige menuer.



Figur 1.1: Et tælletræ over menusammensætning. »F1« er forret nr. 1, »H1« er hovedret nr. 1, osv.

Der er egentlig ikke noget galt i at finde frem til resultatet ved at tegne træer, men opgaven bliver hurtigt uoverskuelig, hvis der er mere end at par enkelte valg, der skal træffes. Man kan i stedet ræsonnere på følgende måde: Der er for hver af de 3 forretter 4 hovedretter at vælge imellem, og for hver af disse er der 2 muligheder for at vælge dessert; det giver i alt

$$3 \cdot 4 \cdot 2 = 24$$

forskellige menuer.

Udregningen i ovenstående eksempel viser det, man kalder »multiplikationsprincippet«. Her er antallet af muligheder ved hvert *delvalg* (forret, hovedret og dessert) blevet ganget med hinanden. Princippet kaldes også *både-og-princippet*, fordi det skal bruges i tilfælde som dette, hvor man skal vælge *både* en forret *og* en hovedret *og* en dessert. Så man har:

Sætning 1.2: Multiplikationsprincippet

Hvis M består af m elementer, og N består af n elementer, kan man vælge et element fra M *og* et element fra N på

$$m \cdot n$$

forskellige måder.

Der findes et andet princip, som kaldes *additionsprincippet* – fordi man her lægger sammen. Dette princip skal bruges i følgende situation.

Eksempel 1.3 En fattig studerende kommer ind på restauranten fra forrige eksempel. Han har kun råd til én ret, så han må vælge mellem enten en forret, en hovedret eller en dessert. I den situation er der

$$3 + 4 + 2 = 9$$

muligheder for at vælge en ret, idet der er 9 forskellige retter i alt, og han kun kan vælge én.

Additionsprincippet kaldes også *enten-eller-princippet*, idet der skal vælges *enten* en forret *eller* en hovedret *eller* en dessert.

Man har altså følgende:

Sætning 1.4: Additionssprincippet

Hvis M består af m elementer, og N består af n elementer, kan man vælge et element fra M *eller* et element fra N på

$$m + n$$

forskellige måder.

1.2 Permutationer

Et andet spørsmål, man kan stille sig selv, handler om at vælge et antal ud af en større mængde. Hvis man f.eks. skal vælge 3 elementer ud af en mængde på 5, hvor mange måder kan det så gøres på?

For at kunne stille en generel formel op, får man brug for noget notation, som gør formlerne nemmere at læse.

Definition 1.5

Hvis n er et naturligt tal, definerer man $n!$,

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

$0!$ defineres til $0! = 1$.
Tallet $n!$ kaldes »n fakultet«.

Eksempel 1.6 Tallet $6!$ er tallet

$$6! = 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 720 \ .$$

Som man måske kan fornemme, kan $n!$ blive et temmeligt stort tal, selv for små værdier af n.

For at finde frem til et svar på det indledende spørgsmål, kan man analysere følgende eksempel:

Eksempel 1.7 Til en filmaften er der 5 forskellige film, der kunne være interessante at se. Men der er kun tid til at se 3 film, så på hvor mange måder, kan man udvælge de tre film, hvis rækkefølgen ikke er ligegyldig?

I dette tilfælde, er der 5 måder at vælge den første film på, 4 måder at vælge den næste (da én film allerede er valgt) og 3 måder at vælge den sidst film på. Det giver i alt

$$5 \cdot 4 \cdot 3 = 60$$

forskellige måder at vælge de tre film på.

Udregningen i eksemplet kan omskrives på følgende måde:

$$5 \cdot 4 \cdot 3 = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1} = \frac{5!}{2!} = \frac{5!}{(5-3)!} \ .$$

Ud fra dette kan man argumentere for følgende generelle formel:

Sætning 1.8

Hvis man skal vælge r elementer ud af n, kan det gøres på $P(n,r)$ måder, hvis rækkefølgen har betydning. Tallet $P(n,r)$ kan udregnes vha. følgende formel:

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

I de tilfælde, hvor $n = r$ (altså hvor man skal finde ud af, hvor mange måder man kan vælge alle elementerne på), taler man om *permutationer*. Antallet af permutationer af en mængde fortæller på hvor mange måder mængden kan sorteres.

Somme tider anvendes benævnelsen *permutationer* om tallet $P(n,r)$, selvom det mere korrekte her ville være at tale om »r-permutation af en n-mængde«.

1.3 Kombinationer

Hvis rækkefølgen ikke betyder noget, er der færre muligheder for at vælge et antal elementer ud af en større mængde. Hvis man skal vælge 3 ud af 5, er problemet ikke større, end at man kan tælle sig frem ved at gå systematisk til værks. Hvis man f.eks. skal vælge tre bogstaver ud af ABCDE kommer man frem til mulighederne i tabel 1.2. Der er altså 10 forskellige måder at vælge 3 ud af 5 på.

Tabel 1.2: Mulighederne for at vælge tre bogstaver ud af ABCDE.

ABC	ACD	BCD	CDE
ABD	ACE	BCE	
ABE	ADE	BDE	

Grunden til, at der bliver færre muligheder, hvis rækkefølgen er ligegyldig, er, at i det tilfælde vil f.eks. ABC og CBA være det samme valg. Når rækkefølgen *har* betydning, kan man vælge 3 ud af 5 på

$$P(5,3) = \frac{5!}{(5-3)!} = 60$$

forskellige måder. Disse 60 muligheder falder i grupper à 6, der indeholder de samme 3 elementer. 3 elementer kan nemlig *permuteres* på $P(3,3) = 5$ forskellige måder. Dvs. hvis rækkefølgen *ikke* har betydning, kan 3 ud af 5 kun vælges på

$$\frac{P(5,3)}{P(3,3)} = \frac{5!}{(5-3)! \cdot 3!} = 10$$

forskellige måder. Denne udregning kan generaliseres til følgende sætning:

Sætning 1.9

Man kan vælge r elementer ud af n på $K(n,r)$ måder, hvis rækkefølgen er uden betydning. Tallet $K(n,r)$ kan beregnes som

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

Tallet $K(n,r)$ kaldes *binomialkoefficienten.*

Der anvendes en del forskellig notation for antallet af kombinationer; foruden $K(n,r)$ ser man også tallet benævnt $K(n,r)$ eller $\binom{n}{r}$.

Eksempel 1.10 I et sæt spillekort er der 52 kort. Hvis man skal vælge 5 af disse kan det gøres på

$$K(52,5) = \frac{52!}{5! \cdot (52-5)!} = \frac{52!}{5! \cdot 47!} = 2\,598\,960$$

forskellige måder.

Sandsynlighedsregning

2

Sandsynlighedsregning handler, som tidligere nævnt, om at kunne sætte tal på tilfældige fænomener. De første bøger om sandsynlighedsregning beskæftigede sig især med diverse spil,[5] og formålet var at finde frem til sandsynlighederne for de enkelte *udfald* af spil.

Et *udfald* kan mere generelt forstås som resultat af et »eksperiment«, dvs. en handling eller en hændelse der kan have flere forskellige resultater.

2.1 Hvad er sandsynlighed?

Kaster man en terning, vil sandsynligheden for at få en 5'er være $\frac{1}{6}$; men hvad betyder det egentligt? Når man kigger på en almindelig sekssidet terning, går man ud fra at der ikke er noget der gør det ene udfald mere sandsynligt end et andet, og man kan så finde sandsynligheden vha. denne formel:

$$\text{sandsynlighed} = \frac{\text{antal gunstige udfald}}{\text{antal mulige udfald}}. \tag{2.1}$$

I dette tilfælde er der tale om såkaldte *a priori* sandsynligheder, dvs. sandsynligheder der er givet på forhånd. Der er tale om en sandsynlighed man tænker sig til ved at analysere eksperimentet.

Man kan dog også bestemme sandsynligheden ved at man slår med terningen et meget stort antal gange og beregner i hvor mange procent af tilfældene man får 5'ere. En sådan sandsynlighed kaldes en *frekventiel* sandsynlighed.

Formlen ovenfor kan som sagt kun bruges når man analyserer en situation hvor alle udfaldene er lige sandsynlige. Det giver derfor god mening at forsøge at beskrive situationer man vil finde sandsynligheden af, på en sådan måde at de udfald man undersøger, er lige sandsynlige.

2.2 Sandsynlighedsfelter

Ser man på et kast med en terning, så er der 6 muligheder for hvad terningen kan vise. Disse muligheder er opsummeret i tabel 2.1. De 6 muligheder er lige sandsynlige, og disse udgør *udfaldsrummet* U, der er mængden af alle mulige udfald,

$$U = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\} \ .$$

Tabel 2.1: Mulige udfald ved et kast med en terning.

u	p
$\boxdot$	$\frac{1}{6}$
$\boxdot$	$\frac{1}{6}$
$\boxdot$	$\frac{1}{6}$
$\boxdot$	$\frac{1}{6}$
$\boxdot$	$\frac{1}{6}$
$\boxdot$	$\frac{1}{6}$

Summen af sandsynlighederne for alle udfald er 1. Udfaldsrummet U og de tilhørende sandsynligheder p kaldes tilsammen for et *sandsynlighedsfelt* (U, p). Formelt har man følgende definition:

Definition 2.1

Hvis $U = \{u_1, \ldots, u_n\}$ er en mængde af udfald, og $p_1, \ldots, p_n$ er de tilhørende sandsynligheder sådan at

1. tallene $p_1, \ldots, p_n$ ligger mellem 0 og 1, og
2. $p_1 + p_2 + \cdots p_n = 1$,

så kaldes (U, p) et endeligt[1] sandsynlighedsfelt.

[1] Der findes også »uendelige« sandsynlighedsfelter hvor antallet af elementer er uendeligt (f.eks. »alle hele tal«), men de ligger uden for rammerne af denne fremstilling.

Hvis man skal beregne sandsynlighederne $p_1, \ldots, p_n$, bliver det nemmere hvis udfaldsrummet er valgt sådan at alle udfald er lige sandsynlige; man taler i dette tilfælde om et *symmetrisk sandsynlighedsfelt*. Det gode ved et symmetrisk sandsynlighedsfelt er at her gælder formel (2.1). Dvs. et symmetrisk sandsynlighedsfelt er defineret på denne måde:

Definition 2.2

Hvis der for et sandsynlighedsfelt (U, p) med n elementer gælder at

$$P(u_1) = P(u_2) = \cdots = P(u_n) = \frac{1}{n},$$

kaldes sandsynlighedsfeltet (U, p) et *symmetrisk* sandsynlighedsfelt.

Enhver delmængde af et udfaldsrum, dvs. en mængde af nogle af udfaldene fra udfaldsrummet, kaldes en *hændelse*. En hændelse er altså en betegnelse for nogle bestemte udfald, som man kigger på i en given situation (det svarer til de »gunstige udfald« i formlen 2.1).

Nogle mulige hændelser ved et kast med en terning kunne være:

$$H_1 = \{\boxdot\}$$
$$H_2 = \{\boxdot, \boxdot\}$$
$$H_3 = \{\boxdot, \boxdot, \boxdot\} \ .$$

Hændelsen H_1 svarer til at få en 6'er, H_2 svarer til at få en 1'er eller en 2'er, mens H_3 svarer til at terningen viser et ulige antal æjne. Alle disse hændelser er altså noget der kan forekomme når man slår en enkelt gang med terningen.

[2] P'et kommer fra det engelske ord *probability*, der betyder »sandsynlighed«.

Til hver hændelse H hører der en sandsynlighed $P(H)$,[2] Man kan beskrive fordelingen af sandsynligheder ved at lave en tabel over sandsynlighederne for hvert enkelt udfald. Sandsynlighederne for et kast med en terning kan ses i tabel 2.1.

Man beregner sandsynligheden for en hændelse ved at lægge sandsynlighederne for alle de udfald der udgør hændelsen, sammen. For de tre hændelser beskrevet ovenfor får man:

$$P(H_1) = p_6 = \tfrac{1}{6}$$

$$P(H_2) = p_1 + p_2 = \tfrac{1}{6} + \tfrac{1}{6} = \tfrac{1}{3}$$
$$P(H_3) = p_1 + p_3 + p_5 = \tfrac{1}{6} + \tfrac{1}{6} + \tfrac{1}{6} = \tfrac{1}{2} .$$

Her er p_1 sandsynligheden for at få en 1'er, osv.

Når man regner på sandsynligheder, kan man nogle gange komme ud for at tage stilling til sandsynligheden for at to (eller flere) ting sker samtidig. Her er det vigtigt at vide om hændelserne er såkaldt *uafhængige* hændelser. Man har følgende:

Definition 2.3

Hvis der for to hændelser A og B i et sandsynlighedsfelt (U, p) gælder at

$$P(\text{både } A \text{ og } B) = P(A) \cdot P(B) ,$$

så kaldes de to hændelser *uafhængige*.

Her følger et eksempel på to hændelser der er uafhængige, og to der ikke er:

Eksempel 2.4 Hvis man kaster en terning to gange, så er de to hændelser

$$A : \text{man får en 6'er i første kast}$$
$$B : \text{man får en 2'er i andet kast}$$

uafhængige hændelser, fordi resultatet af det første kast ikke har nogen indflydelse af resultatet af det andet kast.

Et eksempel på to hændelser der *ikke* er uafhængige har man hvis man fylder 5 sorte og 5 røde bolde i en krukke og trækker to bolde op af krukken. Ser man på de to hændelser

$$C : \text{man trækker en sort bold i første forsøg}$$
$$D : \text{man trækker en rød bold i andet forsøg} ,$$

så er de ikke uafhængige fordi sandsynligheden af det andet udtræk afhænger af om man trak en sort eller en rød bold op første gang.

Det ovenstående bruges når man ser på sandsynligheden for at to hændelser begge indtræffer. Sandsynligheden for at *enten* en hændelse A *eller* en anden hændelse B indtræffer, kan man også beregne i de tilfælde hvor de to hændelser ikke har fælles udfald (dvs. ingen af udfaldene i A må være en del af B og omvendt). I de specielle tilfælde er

$$P(\text{enten } A \text{ eller } B) = P(A) + P(B) .$$

2.3 Stokastiske variable

De tre hændelser H_1, H_2 og H_3 ovenfor kan også beskrives ved såkaldte *stokastiske variable* som knytter et tal til de hændelser man ser på (dvs. hhv. »6'er«, »1'er eller 2'er«, »ulige tal«). Værdierne for to stokastiske variable X og Y kan ses i tabel 2.2.

Tabel 2.2: Udfald ved kast med en terning, samt de to stokastiske variable X og Y.

u	$P(u)$	X	Y
⚀	$\frac{1}{6}$	1	0
⚁	$\frac{1}{6}$	2	1
⚂	$\frac{1}{6}$	3	0
⚃	$\frac{1}{6}$	4	1
⚄	$\frac{1}{6}$	5	0
⚅	$\frac{1}{6}$	6	1

Værdierne af de stokastiske variable er valgt sådan at de beskriver hændelsen bedst muligt – i dette tilfælde ser man på tallene på terningen, eller om de er lige eller ulige. Tallene kan derudover være fuldstændigt tilfældige, men i praksis vil man ofte vælge dem så de beskriver hændelsen bedst muligt. Hændelserne H_1, H_2 og H_3 kan i dette tilfælde skrives som

$$H_1 = \{X = 6\} \ , \quad H_2 = \{X \leq 2\} \ , \quad H_3 = \{Y = 0\} \ .$$

Dvs. sandsynlighederne for de tre hændelser kan skrives på denne måde:

$$P(H_1) = P(X = 6) \ , \quad P(H_2) = P(X \leq 2) \quad \text{og} \quad P(H_3) = P(Y = 0) \ .$$

Ud fra dette eksempel virker stokastiske variable måske som et unødvendigt ekstra lag, for man kan jo bare tælle sig frem. Men så snart situationen bliver bare en anelse mere kompliceret giver begrebet meget mere mening.

Eksempel 2.5 Hvis man kaster to terninger og er interesseret i øjensummen, så kan man skrive et udfaldsrum op bestående af tallene 2, 3, … 12. Disse udfald er bare ikke alle lige sandsynlige idet man f.eks. kun kan få 12 ved at slå to 6'ere, mens man kan få 5 ved både at slå en 1'er og en 4'er eller en 2'er og en 3'er. Men fordi det er nemmere at regne på sandsynligheder hvis sandsynlighedsfeltet er symmetrisk, så analyserer man i stedet situationen på en anden måde:

Tabel 2.3: De mulige værdier for X: summen af øjnene ved et kast med to terninger.

	⚀	⚁	⚂	⚃	⚄	⚅
⚀	2	3	4	5	6	7
⚁	3	4	5	6	7	8
⚂	4	5	6	7	8	9
⚃	5	6	7	8	9	10
⚄	6	7	8	9	10	11
⚅	7	8	9	10	11	12

Hvis man skelner mellem de to terninger ved at lade den ene være hvid og den anden sort, består udfaldsrummet U af par $(h; s)$, hvor h er øjnene på den hvide terning, og s øjnene på den sorte. Der er i alt 36 af sådanne par, udfaldsrummet har altså 36 elementer:

U = { (⚀; **⚀**), (⚀; **⚁**), (⚀; **⚂**), (⚀; **⚃**), (⚀; **⚄**), (⚀; **⚅**),
(⚁; **⚀**), (⚁; **⚁**), (⚁; **⚂**), (⚁; **⚃**), (⚁; **⚄**), (⚁; **⚅**),
(⚂; **⚀**), (⚂; **⚁**), (⚂; **⚂**), (⚂; **⚃**), (⚂; **⚄**), (⚂; **⚅**),
(⚃; **⚀**), (⚃; **⚁**), (⚃; **⚂**), (⚃; **⚃**), (⚃; **⚄**), (⚃; **⚅**),
(⚄; **⚀**), (⚄; **⚁**), (⚄; **⚂**), (⚄; **⚃**), (⚄; **⚄**), (⚄; **⚅**),
(⚅; **⚀**), (⚅; **⚁**), (⚅; **⚂**), (⚅; **⚃**), (⚅; **⚄**), (⚅; **⚅**) }

Alle disse udfald er lige sandsynlige, dvs. sandsynligheden af ét af udfaldene er $\frac{1}{36}$. Sandsynligheden for parret (⚀; **⚅**), svarende til at hvid er en 1'er og sort er en 6'er, er altså $\frac{1}{36}$.

Man kan nu definere den stokastiske variabel X hvis værdi er summen af øjnene på to terninger. De mulige værdier af X er så tallene fra 2 til 12, og vha. tabel 2.3, kan man skrive sandsynlighedsfordelingen for X op.

Tabel 2.4: Sandsynlighedsfordelingen for summen af øjnene ved et kast med to terninger.

t	$P(X = t)$
2	$\frac{1}{36}$
3	$\frac{1}{18}$
4	$\frac{1}{12}$
5	$\frac{1}{9}$
6	$\frac{5}{36}$
7	$\frac{1}{6}$
8	$\frac{5}{36}$
9	$\frac{1}{9}$
10	$\frac{1}{12}$
11	$\frac{1}{18}$
12	$\frac{1}{36}$

Hvis man f.eks. vil vide, hvor sandsynligt det er, at få øjensummen 9 ved et kast med to terninger, kan man tælle antallet af 9-taller i tabel 2.3 og gange med $\frac{1}{36}$, som er sandsynligheden for ét bestemt udfald. Der er fire 9-taller, så

$$P(X = 9) = 4 \cdot \frac{1}{36} = \frac{1}{9} \ .$$

Den samlede sandsynlighedsfordeling for X kan ses i tabel 2.4.

Eksempel 2.6 Hvis man kaster en mønt 3 gange og tæller antallet af »krone«, så er der 4 forskellige muligheder for, hvor mange gange man kan få »krone«: 0, 1, 2 eller 3 gange. Men disse muligheder er ikke lige sandsynlige, og det er derfor ikke smart at betegne dem som udfald.

Man beskriver i stedet udfaldene som de kast, der rent faktisk kan forekomme, dvs. kombinationer af »plat« og »krone«:

$$\text{ppp, ppk, pkk, osv.}$$

Hvis man kigger på antallet af »krone« i 3 kast med en mønt, så er der tre mulige udfald der begge resulterer i 2 gange »krone«. Hændelsen, der udgør disse tre udfald, kan beskrives på denne måde

$$H = \{\text{kkp}, \text{kpk}, \text{pkk}\} \ .$$

Udfaldsrummet består af alle de mulige udfald, dvs.

$$U = \{\text{kkk}, \text{kkp}, \text{kpk}, \text{pkk}, \text{kpp}, \text{pkp}, \text{ppk}, \text{ppp}\} \ .$$

Her ser man at der i alt er 8 mulige udfald som er lige sandsynlige.

Man lader nu den stokastiske variabel X betegne antallet af »krone« i de tre kast, og man kan så lave en tabel over udfaldene og værdierne af den stokastiske variabel (se tabel 2.5). Hændelsen H som blev nævnt ovenfor, svarer så til $X = 2$.

Man kan se i tabellen, at $X = 2$ forekommer 3 steder, dvs.

$$P(X = 2) = 3 \cdot \tfrac{1}{8} = \tfrac{3}{8} \ .$$

Den samlede sandsynlighedsfordeling for antallet af »krone« i 3 kast med en mønt kan ses i tabel 2.6.

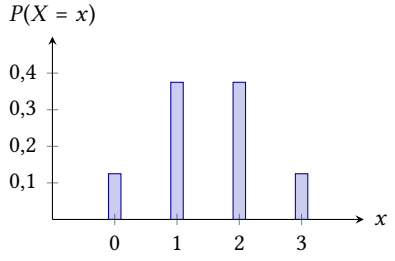
Man kan også præsentere sandsynlighedsfordelingen på er som søjlediagram. Et søjlediagram for sandsynlighedsfordelingen i tabel 2.6 kan ses på figur 2.7.

Tabel 2.5: Værdierne af den stokastiske variabel X for alle mulige udfald af 3 kast med en mønt.

u	X
ppp	0
ppk	1
pkp	1
pkk	2
kpp	1
kpk	2
kkp	2
kkk	3

Tabel 2.6: X's sandsynlighedsfordeling.

x	$P(X = x)$
0	$\frac{1}{8}$
1	$\frac{3}{8}$
2	$\frac{3}{8}$
3	$\frac{1}{8}$



Figur 2.7: Sandsynlighedsfordelingen for X som søjlediagram.

2.4 Middelværdi og spredning

Sandsynlighedsregning kan på mange måder sammenlignes med statistik hvor man i stedet for observationer og frekvenser taler om udfald og sandsynligheder, men metoderne er på mange måder de samme. F.eks. kan sandsynlighedsfordelinger også illustreres vha. søjlediagrammer som det blev gjort i eksemplet i sidste afsnit.

Hvis man har en stokastisk variabel med en tilhørende sandsynlighedsfordeling, kan man derfor også beregne dens middelværdi, varians og spredning.

Definition 2.7

For en stokastisk variabel X der kan antage værdierne $x_1, \ldots, x_n$, er middelværdien μ og spredningen σ givet ved[3]

$$\mu = E(X) = \sum_{i=0}^{n} x_i \cdot P(X = x_i)$$

$$\sigma^2 = \mathrm{Var}(X) = \sum_{i=0}^{n} (x_i - \mu)^2 \cdot P(X = x_i)$$

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

[3] E'et i betegnelsen $E(X)$ komme fra engelsk »expectation value«. Middelværdien kaldes også somme tider den *forventede værdi*.

Man ser her at formlerne er magen til dem der bruges i statistik, men hvor frekvensen f_i er erstattet af sandsynlighederne $P(X = x_i)$.

Eksempel 2.8 Hvis en stokastisk variabel X tæller antallet af »krone« i 3 kast med en mønt, så er dens sandsynlighedsfordeling givet ved tabel 2.6, så er middelværdien

$$\mu = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = 1{,}5\,,$$

variansen er

$$\sigma^2 = (0 - 1{,}5)^2 \cdot \frac{1}{8} + (1 - 1{,}5)^2 \cdot \frac{3}{8} + (2 - 1{,}5)^2 \cdot \frac{3}{8} + (3 - 1{,}5)^2 \cdot \frac{1}{8} = \frac{3}{4}\,,$$

og spredningen er

$$\sigma = \sqrt{\frac{3}{4}} = 0{,}866\,.$$

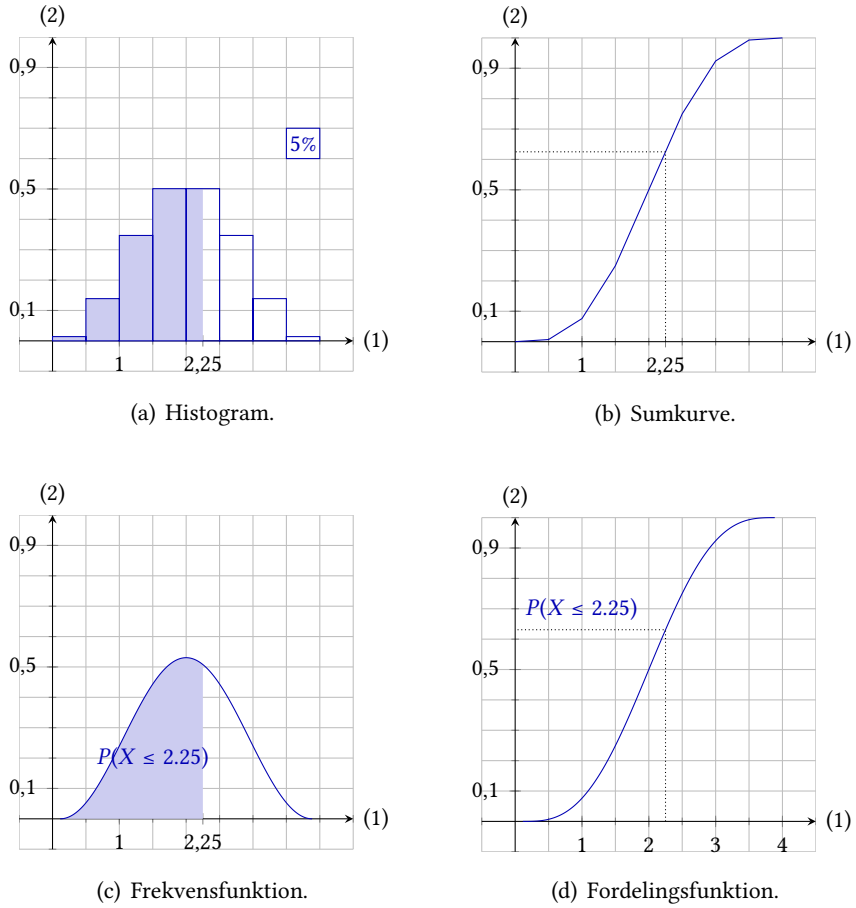
2.5 Diskrete og kontinuerte sandsynligheder

Hvis udfaldsrummet som ved et kast med en terning består af en række adskilte udfald, taler man om et *diskret* sandsynlighedsfelt. Kaster man med en terning, kan man ikke få alle tal mellem 1 og 6, f.eks. kan man ikke få 1,9 eller 2,7. Udfaldsrummet er derfor diskret. Et andet eksempel kunne være et kast med en mønt hvor man tæller antallet af »krone«. Her kan man få hele tal, 1, 2, 3, … – men ikke f.eks. 2,5. De metoder man anvender når man ser på diskrete sandsynligheder, ligner dem man anvender inden for ugrupperet statistik.

Hvis udfaldet derimod kan ramme *alle tal* mellem en minimums- og en maksimumsværdi, så taler man om *kontinuerte* sandsynligheder. Her anvender man metoder der minder om dem der anvendes inden for grupperet statistik. Sandsynlighederne beregnes så i dette tilfælde på intervaller i stedet for på enkelte værdier.

Eksempel 2.9 Hvis der i løbet af et efterårsdøgn måles en minimumstemperatur på 0°C og en maksimumstemperatur på 4°C, så er temperaturen på et tilfældigt tidspunkt i løbet af dette døgn en kontinuert stokastisk variabel, idet temperaturen kan antage alle værdier i intervallet $[0; 4]$.

Hvis man antager, at man har målt temperaturen løbende i løbet af døgnet, kan man fremstille en histogram der viser fordelingen af temperaturer, og



(a) Histogram.

(b) Sumkurve.

(c) Frekvensfunktion.

(d) Fordelingsfunktion.

Figur 2.8: Sammenhængen mellem histogram og sumkurve, og frekvens- og fordelingsfunktion.

en sumkurve der viser i hvor stor en del af døgnet, temperaturen lå under en given værdi. De to diagrammer kan f.eks. se ud som på figur 2.8(a) og 2.8(b).

Har man målt temperaturen løbende, kan man i princippet gøre intervallerne som histogrammet og sumkurven dækker, smallere og smallere. Dvs. hvis man forestiller sig at man gør søjlerne »uendeligt smalle«, vil man til sidst ende med graferne for to kontinuerte funktion, se figur 2.8(c) og 2.8(d).

Disse to funktioner kalder man *frekvensfunktionen* og *fordelingsfunktionen*. Lige som histogrammet arealet under histogrammet angiver frekvensen for en observation, angiver arealet under frekvensfunktionen sandsynligheden for en hændelse, og lige som sumkurven vokser fra 0 til 1, så er fordelingsfunktionen voksende fra 0 til 1.

Som det fremgår af eksemplet, kan kontinuerte stokastiske variable beskrives ved deres frekvensfunktion eller fordelingsfunktion. Sandsynligheden for at et udfald er mindre end en given værdi, kan så findes ved at bestemme arealet under grafen for frekvensfunktionen op til denne værdi. Men den kan også aflæses direkte på fordelingsfunktionen. Hvis fordelingsfunktionen for en kontinuert stokastisk variabel X kaldes F, gælder der

altså

$$P(X \leq a) = F(a) \, .$$

Skal man beregne sandsynligheden for at et udfald ligger i et interval $[a; b]$, får man

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a) \, .$$

Generelt gælder der følgende for kontinuerte sandsynlighedsfordelinger:

Sætning 2.10

Hvis X er en kontinuert stokastisk variabel med fordelingsfunktion F, så er

1. $P(X \leq a) = F(a)$
2. $P(X \geq a) = 1 - F(a)$
3. $P(a \leq X \leq B) = F(b) - F(a)$.

Specielt gælder der for kontinuerte stokastiske variable at sandsynligheden for at få et specifikt udfald er 0 fordi

$$P(X = a) = P(a \leq X \leq a) = F(a) - F(a) = 0 \, .$$

Dette (måske ikke specielt intuitive) resultat skyldes at når en stokastisk variabel er kontinuert, så findes der uendeligt mange tal i udfaldsrummet. Sandsynligheden for at få en specifik værdi bliver derfor uendeligt lille, dvs. 0.

Sammenhængen med integralregning

Idet man finder sandsynlighederne for en hændelse i et kontinuert sandsynlighedsfelt ved at bestemme arealet under en graf, kan denne sandsynlighed bestemmes vha. integration. Kender man til integralregning, kan frekvensfunktioner defineres på følgende måde:

Definition 2.11

En frekvensfunktion f er en kontinuert funktion, hvor $f(x) \geq 0$ for alle $x \in U$, og som opfylder at

$$\int_{-\infty}^{\infty} f(x) \, \mathrm{d}x = 1 \, .$$

Sandsynligheden for en hændelse kan så beregnes på denne måde:

Sætning 2.12

Lad X være en kontinuert stokastisk variabel med frekvensfunktion f. Lad intervallet $H = [h_1; h_2]$ være en hændelse. Da er sandsynligheden $P(H)$ af hændelsen H givet ved

$$P(H) = \int_{h_1}^{h_2} f(x) \, \mathrm{d}x \, .$$

Idet værdien af fordelingsfunktionen er givet ud fra $F(a) = P(X \leq a)$, defineres fordelingsfunktioner på denne måde:

Definition 2.13

Lad X være en kontinuert stokastisk variabel med frekvensfunktion f. Da er fordelingsfunktionen F givet ved

$$F(a) = \int_{-\infty}^{a} f(x)\,\mathrm{d}x\ .$$

For kontinuerte stokastiske variable beregnes middelværdi og spredning også som integraler. Man har denne definition:

Definition 2.14

For en kontinuert stokastisk variabel X med frekvensfunktion f er middelværdien μ, variansen σ^2, og spredningen σ givet ved

$$\begin{aligned}
\mu &= E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\,\mathrm{d}x \\
\sigma^2 &= \mathrm{Var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)\,\mathrm{d}x \\
\sigma &= \sigma(X) = \sqrt{\mathrm{Var}(X)}\ .
\end{aligned}$$

Binomialfordelingen

3

Binomialfordelingen er en sandsynlighedsfordeling som bruges til at beregne hvor stor sandsynligheden er for at få et bestemt antal succeser i en række af eksperimenter. Det kunne f.eks. være hvor stor sandsynligheden er for at få tre 6'ere når man kaster en terning fem gange.

Når man regner på denne situation, tager man udgangspunkt i et eksperiment, man kalder *basiseksperimentet* som udføres et antal gange. Hver gang eksperimentet udføres, er der en sandsynlighed for succes (som man kalder *basissandsynligheden*, p) og en sandsynlighed for fiasko (som er $1 - p$).[1]

[1] Hvis man ikke får succes, så får man fiasko, dvs. sandsynligheden for succes og sandsynligheden for fiasko må tilsammen give 1.

Et eksempel på et basiseksperimentet er et kast med en terning. Det eksperiment udføres fem gange. Hvis succes er at få en 6'er, så er basissandsynligheden $\frac{1}{6}$ (da der er $\frac{1}{6}$ sandsynlighed for at få en 6'er i et kast med en terning). Sandsynligheden for fiasko er så $1 - \frac{1}{6} = \frac{5}{6}$, hvilket svarer til sandsynligheden for at få noget andet end en 6'er.

Hvis man så vil finde sandsynligheden for at få tre 6'ere i de fem kast, så skal man først overveje at de tre 6'ere kan fås på flere måder. Det kunne f.eks. være de første tre kast der gav 6'ere, eller det kunne være de sidste tre. De tre 6'ere kan altså falde på mange forskellige måder. For at man kan sige noget om sandsynligheden, er det derfor nødvendigt først at vide på hvor mange forskellige måder man kan få tre 6'ere i fem kast.

Dette kan beregnes vha. binomialkoefficienten (se sætning 1.9). Tre 6'ere i fem kast kan man få på $K(5,3) = 10$ forskellige måder.

Én af disse består i, at det er de første tre kast der giver 6'ere. Sandsynligheden for, at et kast giver en 6'er er $\frac{1}{6}$. Dette skal ske i de første tre kast. Fjerde og femte kast må ikke give 6'ere, sandsynligheden for dette er $\frac{5}{6}$. Den samlede sandsynlighed for at få først tre 6'ere og dernæst to kast, der giver noget andet, er derfor

$$\overbrace{\frac{1}{6}\cdot\frac{1}{6}\cdot\frac{1}{6}}^{\text{De første tre}}\cdot\overbrace{\frac{5}{6}\cdot\frac{5}{6}}^{\text{De sidste to}} = \left(\frac{1}{6}\right)^3\cdot\left(\frac{5}{6}\right)^2 .$$

Alle de måder, man kan få de tre 6'ere på, må være lige sandsynlige. Hvis man er interesseret i, hvor sandsynligt det er at få tre 6'ere i fem kast (og altså ikke kun at få tre 6'ere i de første tre kast), må denne sandsynlighed ganges med antallet af måder, man kan få de tre 6'ere på, dvs. sandsynligheden bliver

$$10\cdot\left(\frac{1}{6}\right)^3\cdot\left(\frac{5}{6}\right)^2 = \frac{125}{3888} \approx 0{,}0322 . \tag{3.1}$$

Hvis man samler alle deludregninger under ét kan udregningen (3.1) skrives således

$$K(5,3) \cdot \left(\frac{1}{6}\right)^3 \cdot \left(1 - \frac{1}{6}\right)^{5-3} . \tag{3.2}$$

Her anvender man kun de tal, der oprindeligt ses på, nemlig antallet af 6'ere (3), antallet af kast (5) og sandsynligheden for at få en 6'er i et enkelt kast ($\frac{1}{6}$).

3.1 Den generelle formel

Når man skal skrive en generel formel for binomialfordelingen op, indfører man en stokastisk variabel X der tæller antallet af succeser i n eksperimenter. I hver gentagelse af eksperimentet er der en sandsynlighed for succes, som er p.

Man siger så at X er binomialfordelt med *antalsparameter* n og *basissandsynlighed* p, hvilket skrives $X \sim b(n,p)$. Sandsynligheden for r succeser, $P(X = r)$, kan beregnes ved følgende formel der er en generalisering af beregningen (3.2).

Sætning 3.1

Hvis den stokastiske variable X er binomialfordelt med antalsparameter n og basissandsynlighed p, $X \sim b(n,p)$, så er sandsynligheden for r succeser givet ved

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

Eksempel 3.2 Hvad er sandsynligheden for at få præcis fire 1'ere i 15 kast med en terning?

Den stokastiske variabel der tæller antallet af 1'ere i de 15 kast, er binomialfordelt med antalsparameter 15 og basissandsynlighed $\frac{1}{6}$, $X \sim b\left(15, \frac{1}{6}\right)$. Sandsynligheden for at få de fire 1'ere er derfor

$$\begin{aligned} P(X = 4) &= K(15,4) \cdot \left(\frac{1}{6}\right)^4 \cdot \left(1 - \frac{1}{6}\right)^{15-4} \\ &= 1365 \cdot \left(\frac{1}{6}\right)^4 \cdot \left(\frac{5}{6}\right)^{11} = 0{,}1418 . \end{aligned}$$

Der er altså 14,18% sandsynlighed for at få præcis fire 1'ere i 15 kast med en terning.

Eksempel 3.3 På en lille tropeø i stillehavet er der i gennemsnit oversvømmelse om sommeren hvert 4. år. Sandsynligheden for at der kommer en oversvømmelse i løbet af en enkelt sommer er altså $\frac{1}{4}$.

Den stokastiske variabel der tæller antallet af oversvømmelser i løbet af en 5-års periode må være binomialfordelt, $X \sim b\left(5, \frac{1}{4}\right)$. I løbet af en periode på 5 år kan der forekomme alt mellem 0 og 5 oversvømmelser. Den samlede sandsynlighedsfordeling kan man derfor finde ved at beregne $P(X = 0)$, $P(X = 1), \ldots, P(X = 5)$.

Her er f.eks.

$$P(X = 3) = K(5,3) \cdot \left(\frac{1}{4}\right)^3 \cdot \left(\frac{3}{4}\right)^2 = 0{,}0879 \ .$$

Dette er altså sandsynligheden for, at der er oversvømmelser 3 gange i løbet af en 5-års periode. Den samlede fordeling kan ses i tabel 3.1. Et stolpediagram kan ses på figur 3.2.

Som man kan se både i tabellen og på figuren er, at det er mest sandsynligt, at der kommer oversvømmelse i et enkelt år; men man kan også se, at der er en forholdsvist stor sandsynlighed for, at der slet ikke kommer en oversvømmelse i løbet af de 5 år. Til gengæld er det meget lidt sandsynligt (0,0010) at der kommer oversvømmelse i alle 5 år i en 5-års periode.

Hvis man vil vide, hvor stor sandsynligheden er for, at der er højst én oversvømmelse i løbet af de 5 år, skal man udregne

$$P(X \leq 1) = P(X = 0) + P(X = 1) = 0{,}2373 + 0{,}3955 = 0{,}6328 \ .$$

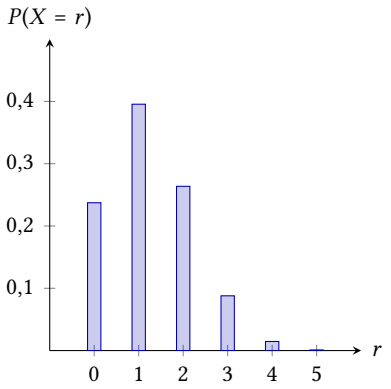
Det er altså meget sandsynligt, at der er højst én oversvømmelse i løbet af de 5 år. Til gengæld er der dog også en sandsynlighed på

$$P(X > 1) = 1 - P(X \leq 1) = 1 - 0{,}6328 = 0{,}3672$$

for at der er mere end 1 oversvømmelse i løbet af 5 år.

Tabel 3.1: Sandsynligheden for, at der er r oversvømmelser i løbet af 5 år.

r	$P(X = r)$
0	0,2373
1	0,3955
2	0,2637
3	0,0879
4	0,0146
5	0,0010



Figur 3.2: Sandsynlighedsfordelingen for X: Antal oversvømmelser i løbet af en 5-års periode.

3.2 Middelværdi og spredning

Middelværdien og spredningen for en binomialfordelt stokastisk variabel er givet ved følgende formler, som ikke bevises:[1]

Sætning 3.4

Hvis den stokastiske variabel X er binomialfordelt, $X \sim b(n, p)$, så er middelværdien μ og spredningen σ givet ved

$$\mu = np$$
$$\sigma = \sqrt{np(1-p)} \ .$$

Eksempel 3.5 Hvis man kaster med en terning 10 gange og tæller antallet af 5'ere, så er den stokastiske variabel, der repræsenterer antallet af 5'ere binomialfordelt med antalsparameter $n = 10$ og basissandsynlighed $p = \frac{1}{6}$.

Middelværdien er så

$$\mu = n \cdot p = 10 \cdot \frac{1}{6} \approx 1{,}667 \ .$$

Hvis man kaster en terning 10 gange, vil man derfor gennemsnitligt få 1,667 5'ere.

Spredningen er

$$\sigma = \sqrt{np(1-p)} = \sqrt{10 \cdot \frac{1}{6} \cdot \left(1 - \frac{1}{6}\right)} = 1{,}179 \ .$$

Eksempel 3.6 I eksempel 3.3 blev der set på en stokastisk variabel med antalsparameter 5 og basissandsynlighed $\frac{1}{4}$.

Her er middelværdien

$$\mu = n \cdot p = 5 \cdot \frac{1}{4} = 1{,}25 \ ,$$

og spredningen er

$$\sigma = \sqrt{np(1-p)} = \sqrt{5 \cdot \frac{1}{4} \cdot \left(1 - \frac{1}{4}\right)} = 0{,}9682 \ .$$

3.3 Binomialtest

James Bond er kendt for at ville have sin martini »shaken not stirred«. Men kan han i virkeligheden smage forskel? Antag at han drikker 16 martinier og skal svare på, hvordan martinien er lavet. I 13 tilfælde svarer han rigtigt. Hvordan kan man afgøre, om det er tilstrækkeligt mange gange til at han ikke bare gætter?

Selve forsøget består 16 gentagelser af basiseksperimentet *bestem om martinien er »shaken« eller »stirred«*. Det er derfor binomialfordelingen der ligger til grund for vurderingen, og det test der skal udføres hedder derfor et *binomialtest*.[2]

[2] I statistik hedder det »et test«, se [3].

Før man udfører testet, skal man opstille en såkaldt *nulhypotese* der kan give en basissandsynlighed. Her vælger man nulhypotesen,

H_0: James Bond kan ikke kende forskel på de to typer martini.

Hypotesen bliver formuleret på denne måde fordi det er det man kan teste. Man ved nemlig hvad sandsynlighederne er hvis han bare gætter (så er sandsynligheden nemlig $\frac{1}{2}$) – derimod kan man ikke sige noget om sandsynlighederne hvis han rent faktisk kan kende forskel.

Herefter vælger man et såkaldt *signifikansniveau* der sætter grænsen for hvornår man vælger at forkaste nulhypotesen. Et typisk valg er 5%. Testet går herefter ud på at undersøge hvor langt resultatet i en stikprøve skal ligge fra middelværdien for at sandsynligheden for resultatet er mindre end signifikansniveauet.

I dette tilfælde vælger man at undersøge sandsynligheder af typen

$$P(X \geq k) \ ,$$

hvor k er et helt tal. Her har man f.eks.

$$P(X \geq 11) = 0{,}105 = 10{,}5\%$$
$$P(X \geq 12) = 0{,}038 = 3{,}8\% \ .$$

Her kan man se at der er mindre end 5% sandsynlighed for at få en værdi der er 12 eller mere. Den såkaldt *kritiske mængde* er derfor

$$\{12, 13, 14, 16\} \ .$$

Denne mængde beskriver de udfald som der tilsammen er mindre end 5% sandsynlighed for, dvs. mindre sandsynlighed end det valgte signifikansniveau.

Idet James Bonds resultat falder i den kritiske mængde, er der altså mindre end 5% sandsynlighed for at se dette resultat hvis han bare gætter. Man vælger derfor at *forkaste* nulhypotesen. Han kan altså godt smage på hvilken måde martinien er lavet.

Testet der blev udført, er et såkaldt *højresidet* test idet man tester om værdien er for stor til at han bare gætter. I andre situationer kan man komme ud for at skulle teste om en given værdi er for lille, og her anvender man en *venstresidet* test:

Højresidet test I et højresidet binomialtest med signifikansniveau α er den kritiske mængde

$$K = \{k, k+1, \ldots, n\} \ ,$$

hvor k er det mindste tal, sådan at $P(X \leq k) \leq \alpha$.

Venstresidet test I et venstresidet binomialtest med signifikansniveau α er den kritiske mængde

$$K = \{0, 1, \ldots, k\} \ ,$$

hvor k er det største tal, sådan at $P(X \leq k) \leq \alpha$.

Eksempel 3.7 Et firma der producerer dåsetomater, lover at 98% af dåserne kommer uskadte frem ved levering. Et supermarked modtager en palle tomater med 950 dåser. De 25 af dåserne har taget skade.

For at finde ud af om man kan stole på firmaets løfter kan man lave et venstresidet binomialtest. Nulhypotesen er i dette tilfælde

$$H_0\text{: 98\% af dåserne er uskadte.}$$

Binomialfordelingen har så $n = 950$ og $p = 98\%$. Med et signifikansniveau på 5%, finder man

$$P(X \leq 923) = 0{,}0468 = 4{,}68\%$$
$$P(X \leq 924) = 0{,}0711 = 7{,}11\% \ .$$

Den kritiske mængde er altså

$$K = \{1, 2, 3, \ldots, 921, 922, 923\} \ .$$

925 af dåserne er uskadte, og dette tal ligger uden for den kritiske mængde, så derfor accepteres nulhypotesen. På et 5% signifikansniveau kan man altså stole på firmaets løfter.

I det test der blev udført ovenover, testede man om James Bond kan kende forskel på to slags martinier ved at se på om antallet af rigtige svar var højt nok til at forkaste nulhypotesen.

Man kan også vælge at lave et *tosidet* test. Her tester man om tallet enten er for lille eller for stort. Hvis man forkaster nulhypotesen her, vil konklusionen være at han kan smage forskel (men ikke at han kan fortælle hvilken drink der er hvad).

Da testet er tosidet, deler man signifikansniveauet med 2 (her bliver det 2,5%) og undersøger hvornår hhv. $P(X \leq a)$ og $P(X \geq b)$ bliver større end 2,5%. Her har man

$$P(X \leq 3) = 0{,}0106 = 1{,}06\%$$
$$P(X \leq 4) = 0{,}0384 = 3{,}84\%$$

og

$$P(X \geq 12) = 0{,}0384 = 3{,}84\%$$
$$P(X \geq 13) = 0{,}0106 = 1{,}06\% \ .$$

I dette tilfælde bliver den kritiske mængde

$$K = \{0, 1, 2, 3\} \cup \{13, 14, 15, 16\} \ .$$

Generelt finder man den kritiske mængde i et tosidet binomialtest på denne måde:

Tosidet test I et tosidet binomialtest med signifikansniveau α er den kritiske mængde

$$K = \{0, 1, \ldots, k\} \cup \{l, \ldots, n\} \ ,$$

hvor k er det største tal, sådan at $P(X \leq k) \leq \frac{\alpha}{2}$, og l er det mindste tal sådan at $P(X \geq l) \leq \frac{\alpha}{2}$.

Eksempel 3.8 I Vaffelbjerg Kommune fik Protestpartiet ved sidste kommunalvalg 17,2% af stemmerne. Ved en meningsmåling hvor man har spurgt 1000 repræsentativt udvalgte personer, siger 243 personer at de vil stemme på partiet ved næste kommunalvalg.

Hvis man vil vide om partiets stemmeandel har ændret sig, kan man udføre et tosidet binomialtest. Nulhypotesen bliver i dette tilfælde

$$H_0\text{: Partiets stemmeandel er 17,2\%.}$$

Binomialfordelingen har så $n = 1000$ og $p = 17{,}2\%$. Hvis signifikansniveauet er $\alpha = 5\%$, så er $\frac{\alpha}{2} = 2{,}5\%$. For den nedre grænse finder man

$$P(X \leq 148) = 0{,}0229 = 2{,}29\%$$
$$P(X \leq 149) = 0{,}028 \qquad\qquad = 2{,}8\%$$

og for den øvre grænse

$$P(X \geq 196) = 0{,}0259 = 2{,}59\%$$
$$P(X \geq 197) = 0{,}0214 = 2{,}14\% \ .$$

Ud fra disse beregninger kan man se at den kritiske mængde er

$$K = \{0, 1, \ldots, 148\} \cup \{197, 198, \ldots, 1000\} \ .$$

Idet tallet 243 ligger i den kritiske mængde, forkaster man nulhypotesen. Partiets vælgertilslutning *har* altså ændret sig siden valget.

Normalfordelingen

4

Mange statistiske målinger resulterer i frekvensfordelinger, der med god tilnærmelse følger den sandsynlighedsfordeling der kaldes normalfordelingen. Et eksempel på dette kunne være tykkelsen af brød skåret på en maskine.

Ingen maskine skærer helt perfekt. I tabel 4.1 ses målinger foretaget på en maskine der skal skære brød i 1 cm tykke skiver. Nogle af skiverne bliver for tykke og nogle for tynde; men det ser dog ud til at de fleste har en tykkelse på omkring 1 cm.

Ud fra tallene i tabellen kan man også finde middelværdien μ og stikprøvespredningen s for tykkelsen af brødskiverne. Det viser sig, at

$$\overline{x} = 1{,}151 \qquad \text{og} \qquad s = 0{,}249\ .$$

På figur 4.2 er der tegnet et histogram over fordelingen fra tabel 4.1. På figuren er også indtegnet grafen for frekvenssfunktionen for *normalfordelingen med middelværdi 1,151 og spredning 0,249*. frekvensfunktionens graf er en klokkeformet kurve, der ser ud til at passe ganske godt med de målte værdier af brødskivernes tykkelse.

Hvis en kontinuert stokastisk variabel X er normalfordelt med middelværdi μ og spredning σ, skriver man $X \sim N(\mu, \sigma)$. Er X normalfordelt, har den følgende frekvensfunktion:

Tabel 4.1: Målinger på 100 skiver brød skåret på en maskine.

Tykkelse (cm)	Antal
0,55–0,65	2
0,65–0,75	4
0,75–0,85	6
0,85–0,95	9
0,95–1,05	14
1,05–1,15	16
1,15–1,25	15
1,25–1,35	10
1,35–1,45	10
1,45–1,55	8
1,55–1,65	6

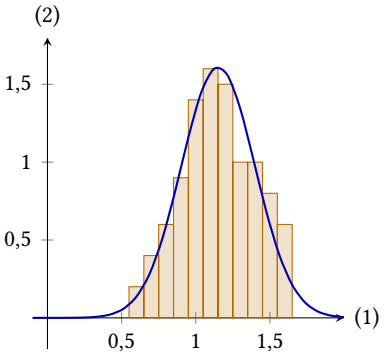
Definition 4.1

En stokastisk variabel X kaldes normalfordelt med middelværdi μ og spredning σ, $X \sim N(\mu, \sigma)$ hvis den har frekvensfunktionen

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

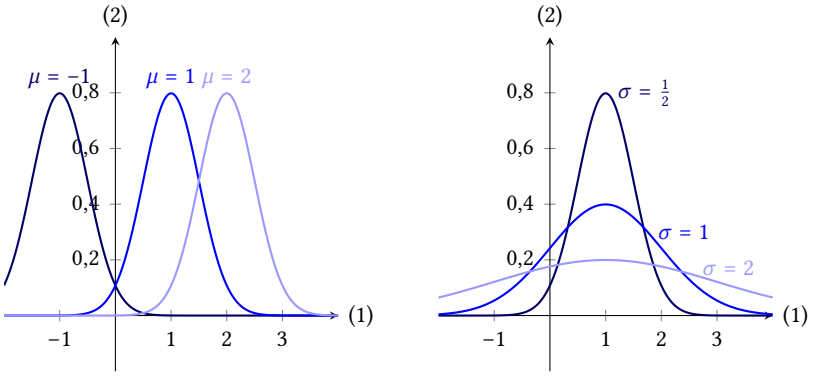
Der er mange fænomener, der resulterer i normalfordelte data. Nogle af dem kunne være

- Målefejl i eksperimenter.
- Størrelsen af ting, der er maskinelt fremstillet (som f.eks. tykkelsen af brødskiverne ovenfor).
- Biologiske variable som f.eks. højde og vægt.[1]



Figur 4.2: Histogram over tykkelsen af brødskiver.

[1] Mange biologiske variable er dog kun tilnærmelsesvis normalfordelte og er i virkeligheden log-normalfordelte.[6].



(a) Forskellig middelværdi.

(b) Forskellig spredning.

Figur 4.3: Hvis frekvensfunktionerne har den samme spredning, men forskellig middelværdi, er graferne forskudt vandret. Har de den samme middelværdi, men forskellig spredning, ændres bredden af kurven.

Middelværdien og spredningen påvirker udseendet af kurven. Hvis man ændrer middelværdien flytter kurven sig i vandret retning. Hvis spredningen bliver mindre, bliver kurven smallere; bliver spredningen større, bliver kurven bredere (se figur 4.3).

Som for alle andre kontinuerte sandsynlighedsfordelinger finder man sandsynligheden for at få en måling i et bestemt interval ved at bestemme arealet under grafen for frekvensfunktionen i det pågældende interval.

Eksempel 4.2 En maskine på en fabrik fylder 1 kg poser med sukker. Vægten af poserne er normalfordelt med middelværdien $\mu = 1000$ g og spredning $\sigma = 25$ g. Frekvensfunktionen er så

$$f(x) = \frac{1}{\sqrt{2\pi} \cdot 25} \cdot \mathrm{e}^{-\frac{(x-1000)^2}{2 \cdot 25^2}} .$$

[2]Det er ikke muligt at skrive et algebraisk udtryk op for fordelingsfunktionen, men den er heldigvis programmeret ind i de fleste CAS-værktøjer

Sandsynligheder beregnes dog nemmest ud fra fordelingsfunktionen, F.[2] Sandsynligheden for i en stikprøve at få en pose, der vejer mellem 950 og 975 g bliver

$$\begin{aligned} P(950 \leq X \leq 975) &= F(975) - F(950) \\ &= 0{,}1359 = 13{,}59\% . \end{aligned}$$

Der er altså en ikke ubetydelig sandsynlighed for at få fat i en pose der vejer et lille stykke under 1 kg.

Eksempel 4.3 Her ses igen på eksemplet med brødskiverne fra starten af kapitlet. Det viste sig, at den stokastiske variabel der angiver tykkelsen af brødskiverne med god tilnærmelse er normalfordelt med middelværdi $\mu = 1{,}151$ og spredning $\sigma = 0{,}249$.

Ud fra dette kan man få svar på spørgsmål som

1. Hvad er sandsynligheden for at få en skive brød, hvis tykkelse er mellem 0,9 cm og 1 cm?
2. Hvad er sandsynligheden for at få en skive brød, hvis tykkelse er over 1,3 cm?

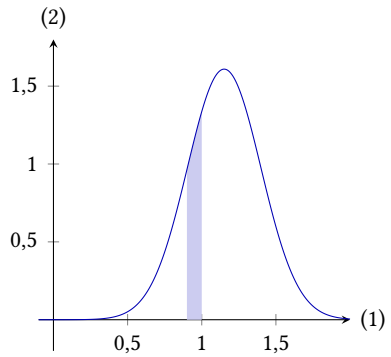
Begge svar findes nemmest ved at bruge fordelingsfunktionen til at udregne arealet under grafen for frekvensfunktionen. Sandsynligheden for at få en skive brød med en tykkelse mellem 0,9 cm og 1 cm er så

$$P(0{,}9 \le X \le 1) = F(1) - F(0{,}9) = 0{,}1154 \,.$$

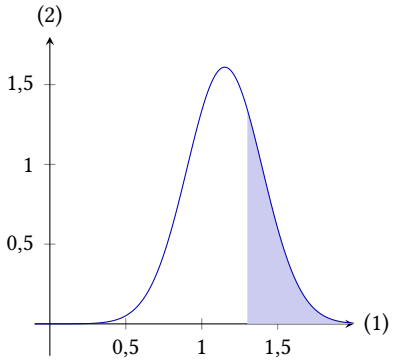
Sandsynligheden for at få en skive brød, der er tykkere end 1,3 cm er

$$P(X \ge 1{,}3) = 1 - P(X \le 1{,}3) = 1 - F(1{,}3) = 0{,}2748 \,.$$

De arealer, der repræsenterer sandsynlighederne kan ses på figur 4.4.



(a) $P(0{,}9 \le X \le 1) = 0{,}1154$.



(b) $P(X \ge 1{,}3) = 0{,}2748$.

Figur 4.4: Sandsynlighederne for at brødskivernes tykkelser falder i bestemte intervaller kan aflæses som arealet under grafen for tæthedsfunktionen.

Grafen for normalfordelingens frekvensfunktion er symmetrisk omkring middelværdien. Faktisk opfører frekvensfunktionen sig så pænt, at man har følgende sætning som ikke bevises her.[4]

Sætning 4.4

Hvis X er en normalfordelt stokastisk variabel med middelværdi μ og spredning σ er

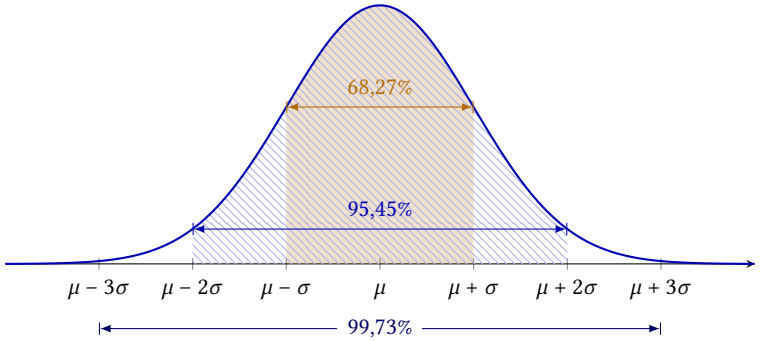
$$P(\mu - \sigma \le X \le \mu + \sigma) = 0{,}6827$$
$$P(\mu - 2\sigma \le X \le \mu + 2\sigma) = 0{,}9545$$
$$P(\mu - 3\sigma \le X \le \mu + 3\sigma) = 0{,}9973 \,.$$

Ud fra denne sætning kan man se, at 68,27% af alle målinger vil ligge i et interval der dækker 1 spredning til hver side af middelværdien. 95,45% vil ligge i et interval på 2 spredninger til hver side af middelværdien, osv. Dette kan ses illustreret på figur 4.5.

Idet langt de fleste udfald (95,45%) ligger inden for intervallet $\mu \pm 2\sigma$, kalder man udfald der ligger i dette interval for *normale* udfald. Udfald der ligger mere end 3σ fra middelværdien, kaldes *eksceptionelle* udfald. Disse udfald udgør nemlig kun $1 - 0{,}9973 = 0{,}0027 = 0{,}27\%$ af den samlede fordeling.

4.1 Approksimation til binomialfordelingen

Hvis man har en stokastisk variabel som er binomialfordelt $X \sim b(n, p)$, viser det sig at man kan approksimere fordelingen af X med en normalfordeling der har samme middelværdi og spredning som binomialfordelingen.



Figur 4.5: For en normalfordelt stokastisk variabel gælder, at sandsynligheden for at X ligger i et symmetrisk interval på 1 spredning til hver side af middelværdien er et fast tal. Det samme gælder for et interval på 2 spredninger til hver side af middelværdien, osv.

Approksimationen er rigtig god når $\sigma^2 = np(1-p) \geq 10$, og den bliver bedre jo større denne størrelse er.[1]

Eksempel 4.5 Hvis en stokastisk variabel $X \sim b(10; 0{,}4)$ er binomialfordelt, er dens middelværdi og spredning

$$\mu = n \cdot p = 10 \cdot 0{,}4 = 4$$
$$\sigma = \sqrt{np(1-p)} = \sqrt{10 \cdot 0{,}4 \cdot 0{,}6} = \sqrt{2{,}4} = 1{,}55\ .$$

σ^2 giver her 2,4 som er en del mindre end 10, dvs. man forventer ikke at normalfordelingen er en specielt god approksimation til denne binomialfordeling.

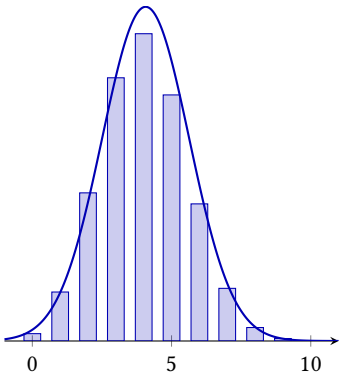
På figur 4.5 ses et søjlediagram over sandsynlighedsfordelingen for X sammen med grafen for frekvensfunktionen for normalfordelingen med middelværdi 4 og spredning 1,55, og som man kan se, er approksimationen nogenlunde, men ikke specielt god.

Ser man derimod på en stokastisk variabel Y der er binomialfordelt $Y \sim b(50; 0{,}4)$, bliver billedet et andet. Her er middelværdien og spredningen

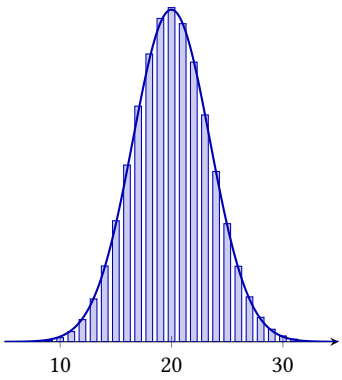
$$\mu = 50 \cdot 0{,}4 = 20$$
$$\sigma = \sqrt{50 \cdot 0{,}4 \cdot 0{,}6} = \sqrt{12} = 3{,}46\ ,$$

og σ^2 er altså 12, dvs. lidt mere end 10. Tegner man nu et søjlediagram over sandsynlighedsfordelingen for Y og grafen for frekvensfunktionen for normalfordelingen med middelværdi 20 og spredning 3,46, får man billedet på figur 4.5.

Her ses en tydelig overensstemmelse mellem grafen og søjlediagrammet, dvs. her er normalfordelingen en god tilnærmelse til binomialfordelingen.



Figur 4.6: Fordelingen $b(10; 0{,}4)$ approksimeret med en normalfordeling.



Figur 4.7: Fordelingen $b(50; 0{,}4)$ approksimeret med en normalfordeling.

4.2 Stikprøver

Idet normalfordelingen anvendes som approksimation til binomialfordelingen, kan den også bruges til at bestemme et *konfidensinterval* for sandsynlighedsparameteren i en binomialfordeling. For at kunne bestemme konfidensintervallet, skal man dog først have et estimat for sandsynlighedsparameteren.

Eksempel 4.6 Et firma vil undersøge hvor mange procent af befolkningen der kender til et nyt produkt de har sendt på markedet. Et analysefirma spørger en stikprøve på 1142 repræsentativt udvalgte mennesker om de kender til det nye produkt. Det viser sig at 715 har hørt om det nye produkt.

Ud fra disse tal, kan man estimere at

$$\hat{p} = \frac{715}{1142} = 0{,}626 = 62{,}6\%$$

af befolkningen har hørt om det nye produkt.

Spørgsmålet er nu hvor sikker man kan være på dette tal? Hvis man antager at $\hat{p}$ er et godt estimat for den sande sandsynlighedsparameter p,

så er antallet af mennesker der har hørt om produktet, binomialfordelt med antalsparameter $n = 1142$ og sandsynlighedsparameter $\hat{p} = 0{,}626$. Middelværdien for denne binomialfordeling er

$$\mu = n\hat{p}\,,$$

og spredningen er

$$\sigma = \sqrt{n\hat{p}(1-\hat{p})}\,.$$

Hvis man så approksimerer denne binomialfordeling med en normalfordeling, finder man at 95,45% af fordelingen vil ligge i intervallet

$$\mu \pm 2\sigma = n\hat{p} \pm 2 \cdot \sqrt{n\hat{p}(1-\hat{p})}\,.$$

Idet man er mere interesseret i den procentvise andel end af antallet, kan man dividere dette tal med n, hvorved man får intervallet

$$\hat{p} \pm 2 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\,.$$

Man kan nu sige at med 95,45% sandsynlighed vil den rigtige værdi for sandsynlighedsparameteren p ligge i dette interval.

Sætning 4.7

Ved en stikprøve med n elementer og n_s succeser, er 95%-konfidensintervallet givet ved

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

hvor $\hat{p} = \frac{n_s}{n}$.

I virkeligheden finder man 95,45%-konfidensintervallet for p, når man bruger denne formel. Hvis man er interesseret i præcis 95%, skal 2-tallet i formlen erstattes med 1,96.[4]

Eksempel 4.8 I eksemplet fra før, blev sandsynlighedsparameteren estimeret til $\hat{p} = 0{,}626$, dvs.

$$2 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 2 \cdot \sqrt{\frac{0{,}626 \cdot (1-0{,}626)}{1142}} = 0{,}029\,.$$

Sandsynlighedsparameteren ligger derfor med 95% sandsynlighed i intervallet

$$0{,}629 \pm 0{,}029$$

hvilket vil sige at med 95% sandsynlighed har mellem 59,7% og 65,5% af befolkningen hørt om det nye produkt.

4.3 Standardnormalfordelingen

Hvis man skal regne på normalfordelingen vil man normalt anvende et CAS-værktøj. Før dette var muligt var man nødt til at anvende tabelopslag.

Da man ikke kan lave en tabel for hver mulig middelværdi og spredning, lavede man i stedet tabeller over den såkaldte *standardnormalfordeling* og udnyttede, at enhver normalfordelings frekvens- og fordelingsfunktion kan udtrykkes vha. standardnormalfordelingen.

Standardnormalfordelingen er defineret på følgende måde.

Definition 4.9

Normalfordelingen med middelværdi $\mu = 0$ og spredning $\sigma = 1$ kaldes *standardnormalfordelingen*. Dens frekvensfunktion er

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

Fordelingsfunktionen for standardnormalfordelingen er

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)\,\mathrm{d}t .$$

I de næste par sætninger vises sammenhængen mellem frekvens- og fordelingsfunktionen for en vilkårlig normalfordeling og de tilsvarende funktioner for standardnormalfordelingen.

Sætning 4.10

Der er følgende sammenhæng mellem frekvensfunktionen $f(x)$ for normalfordelingen med middelværdi μ og spredning σ og frekvensfunktionen for standardnormalfordelingen:

$$f(x) = \frac{1}{\sigma} \cdot \phi\left(\frac{x-\mu}{\sigma}\right) .$$

Bevis
Sætningen bevises ved at regne på $\frac{1}{\sigma} \cdot \phi\left(\frac{x-\mu}{\sigma}\right)$:

$$\begin{aligned}\frac{1}{\sigma} \cdot \phi\left(\frac{x-\mu}{\sigma}\right) &= \frac{1}{\sigma} \cdot \frac{1}{\sqrt{2\pi}} \cdot \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \\ &= \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} = f(x) .\end{aligned}$$

■

Den funktion man fandt i tabeller var dog ikke frekvensfunktionen, men fordelingsfunktionen, idet denne kan bruges direkte til at beregne sandsynligheder. Her gælder følgende sammenhæng:

Sætning 4.11

Sammenhængen mellem fordelingsfunktionen $F(x)$ og standardnormalfordelingens fordelingsfunktion $\Phi(x)$ er

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) .$$

Bevis
Fordelingsfunktionen for normalfordelingen med middelværdien μ og spredningen σ er ifølge sætning 4.10

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = \int_{-\infty}^{x} \frac{1}{\sigma} \cdot \phi\left(\frac{t-\mu}{\sigma}\right)\,\mathrm{d}t . \tag{4.1}$$

Man laver nu substitutionen $u = \frac{t-\mu}{\sigma}$. Herved bliver $\mathrm{d}u = \frac{1}{\sigma}\cdot\mathrm{d}x$, og udtrykket i (4.1) bliver til

$$\int_{-\infty}^{\frac{x-\mu}{\sigma}} \phi(u)\,\mathrm{d}u = \Phi\left(\frac{x-\mu}{\sigma}\right) .$$

Altså er

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) . \qquad \blacksquare$$

Det er ikke nødvendigt at vide dette, hvis man blot skal regne direkte på en given normalfordeling. Men det viser sig, at standardnormalfordelingen kan være nyttig, hvis man vil undersøge, om et givent datasæt er normalfordelt.

4.4 Normalfordelte data

I dette afsnit vises hvordan man kan konstruere et såkaldt *fraktilplot* eller *QQ-plot* som er en graf der kan bruges til at afgøre om et datasæt er normalfordelt.

Eksempel 4.12 Det skal undersøges om følgende talsæt er normalfordelt (tallene viser kontrolvejninger af 30 sække gulerødder der hver skal veje 25 kg):

24,8	25,4	25,0	25,4	24,0	24,4
24,5	24,5	24,8	24,9	24,9	25,0
24,7	24,3	24,7	24,8	25,0	25,1
25,1	24,5	24,3	25,0	25,3	25,1
24,5	24,7	25,3	24,6	24,5	24,8

Hvis man beregner middelværdien og stikprøvespredningen af dette datasæt, får man

$$\overline{x} = 24{,}8 \quad \text{og} \quad s = 0{,}35 .$$

Det man nu gerne vil vide, er om datasættet svarer til en normalfordeling med denne middelværdi og spredning. Først sorterer man de 30 tal og nummererer dem (se tabel 4.8).

Ideen i et fraktilplot er at sammenligne de kumulerede frekvenser med de kumulerede frekvenser for standardnormalfordelingen. Hvis dataene er

Tabel 4.8: Vægten af sække med gulerødder, sorteret og nummereret.

Vægt, x	i	$z = \Phi^{-1}\left(\frac{i-0,5}{n}\right)$
24,0	1	−2,13
24,3	2	−1.64
24,3	3	−1,38
24,4	4	−1,19
⋮	⋮	⋮
25,3	28	1,38
25,4	29	1,64
25,4	30	2,13

normalfordelte så vil de kumulerede være givet ved en fordelingsfunktion $F(x)$ for en normalfordeling med middelværdi $\overline{x}$ og spredning s hvor

$$F(x) = \Phi\left(\frac{x - \overline{x}}{s}\right) ,$$

hvilket kan omskrives til[3]

$$\Phi^{-1}(F) = \frac{x - \overline{x}}{s} .$$

[3]Her er Φ^{-1} den inverse funktion til fordelingsfunktionen for standardnormalfordelingen. Denne funktion findes i de fleste CAS-værktøjer.

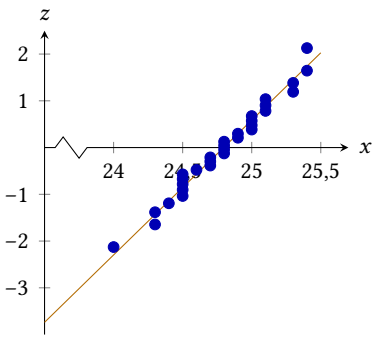
Det vil sige, at $\Phi^{-1}(F)$ er en lineær funktion af x.

Her er F de kumulerede frekvenser. I dette eksempel er de kumulerede frekvenser $F = \frac{i}{30}$ hvor i er målingens nummer. Den første måling har så den kumulerede frekvens $\frac{1}{30}$, mens den sidste har den kumulerede frekvens $\frac{30}{30} = 1$. Dette giver et problem, idet funktionen $\Phi^{-1}(F)$ ikke er defineret for $F = 1$.[4] For at få det sidste punkt med anvender man derfor tallet $\frac{i-0,5}{n}$ (hvor n er antallet af målinger, dvs. her er $n = 30$) stedet for de kumulerede frekvenser[2] og beregner z som

[4]Dette skyldes at fordelingsfunktionen for en normalfordeling aldrig vil antage værdien 1, men kun vil nærme sig 1, når $x \longrightarrow \infty$.

$$z = \Phi^{-1}\left(\frac{i - 0,5}{n}\right) .$$

Herefter afbilder man z som funktion af x, se figur 4.9.



Figur 4.9: Kvartilplottet for vægten af gulerødder.

Hvis datasættet med god tilnærmelse er normalfordelt, så skal punkterne ligge tilnærmelsesvis på den rette linje $z = \frac{x-\overline{x}}{s}$ som i dette tilfælde er

$$z = \frac{x - 24,8}{0,35} .$$

På figuren kan man se at punkterne med god tilnærmelse ligger på denne rette linje, så man kan konkludere at datasættet er normalfordelt.

En sammenfatning af metoden ser således ud:

1. Beregn middelværdien $\overline{x}$ og stikprøvespredningen s.
2. Lav en tabel over datasættet hvori tallene er sorteret og nummereret.
3. Tilføj en kolonne hvori $z = \Phi^{-1}\left(\frac{i-0,5}{n}\right)$.
4. Afsæt punkterne (x, z) i et koordinatsystem. x er målingen.
5. Tegn også linjen $z = \frac{x-\overline{x}}{s}$.
6. Hvis datasættet er normalfordelt, så skal punkterne med god tilnærmelse ligge på denne linje.

Da det er noget besværligt at konstruere fraktilplot manuelt, er det heldigt at de er bygget ind i mange CAS-værktøjer.

Bibliografi

[1] Richard F. Bass m.fl. *Upper level undergraduate probability with actuarial and financial applications*. University of Connecticut, Department of Mathematics, 2018.

[2] Nicholas J. Cox. *Calculating percentile ranks or plotting positions*. URL: https://www.stata.com/ (bes. 23.03.2020).

[3] Claus Thorn Ekstrøm. *Selfølgelig hedder det et test*. 27. jun. 2017. URL: https://sandsynligvis.dk/2017/06/27/selvf%C3%B8lgelig-hedder-det-et-test/ (bes. 23.03.2020).

[4] Sue Gordon. *The Normal Distribution*. Mathematics Learning Centre, 2006.

[5] Victor J. Katz. *A History of Mathematics. An Introduction*. 2. udg. Addison-Wesley Educational Publishers, Inc., 1998.

[6] Eckhard Limpert, Werner A. Stahel og Markus Abbt. »Log-normal Distributions across the Sciences: Keys and Clues«. I: *BioScience* 51.5 (2001), s. 341–352.