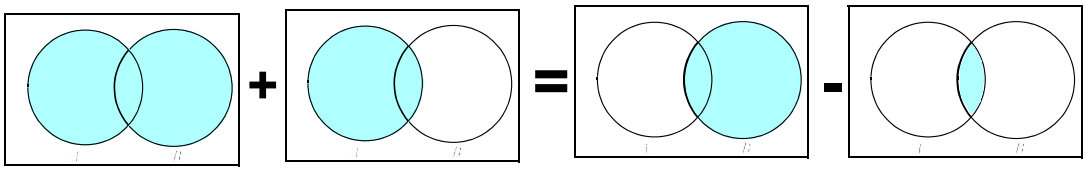
MOGENS ODDERSHEDE LARSEN

Sandsynlighedregning



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

1. udgave 2007

FORORD

Dette notat giver en kort gennemgang af de grundlæggende begreber i sandsynlighedsregning. Det forudsættes, at man har rådighed over en "matematiklommeregner" som eksempelvis Ti 89.

En række noter indenfor matematik og statistik kan findes på adressen www.larsen-net.dk

Nogle af noterne er beregnet for det gymnasiale niveau, andre er beregnet på et mere videregående niveau.

Noterne er i pdf - format.

Maj 2007 Mogens Oddershede Larsen

INDHOLD

1. Definitioner og regneregler

1.1 Indledning

Tilfældigt eksperiment (engelsk : random experiment)
Ved et "tilfældigt eksperiment forstås et eksperiment, som kan resultere i forskellige udfald, selv om eksperimentet gentages på samme måde hver gang. Man kan ikke på forhånd forudsige, hvilket udfald der vil indtræffe.

Eksempler på tilfældige eksperimenter

1) Består eksperimentet i kast med en terning ved vi, at vi vil få et af udfaldene 1,2,3,4,5,6 (Udfaldsrummet $U = \{1, 2, 3, 4, 5, 6\}$), men man kan ikke forudsige udfaldet

2) Består eksperimentet i, at vi som i eksempel 1.4 tilfældigt udtager en patient og måler pH værdien i hans knæ, ved vi måske at værdien vil ligge mellem 5 og 10 (Udfaldsrummet $U = [5;10]$), men vi kan ikke forudsige resultatet.

3) Består eksperimentet i, at vi tilfældigt udtrækker en vælger, og spørger hvilket parti vedkommende vil stemme på hvis der var valg i morgen, så er udfaldsrummet de forskellige opstillingsberettigede partier.

En delmængde af udfaldsrummet kaldes en **hændelse**.
Eksempel: A: at få et lige øjental ved kast med en terning

1.2. Sandsynlighed

Det er en erfaring, at øges antallet af gentagelser af et eksperiment, vil den relative hyppighed af en hændelse A stabilisere sig mod en bestemt værdi.("**de store tals lov**"), som så kaldes "sandsynligheden for A og benævnes $\boldsymbol{P(A)}$ (P = probability) .

Eksempel 1.1. De relative hyppigheders stabilitet

Et eksperiment består i at kaste en terning, og hændelsen A består i at få et lige øjental. Terningen kastes nu 100 gange, og man får et lige øjental 55 gange. Eksperimentet udføres igen 100 gange, og man får A 47 gange. Igen kastes 100 gange, og man får nu A 57 gange. Til sidst kastes 100 gange, og man får A 40 gange.
Eksperimentet foretages nu i serier på 1000 gange, hvor man hver gang optæller antal gange A forekommer. Resultaterne vises i følgende tabel:

	Serier på 100 gentagelser				Serier med 1000 gentagelser			
	1	2	3	4	1	2	3	4
Antal gange A: et lige øjental	55	47	51	40	486	508	488	509
Relativ hyppighed	0.55	0.47	0.51	0.40	0.486	0.508	0.488	0.509

Det ses, at med 1000 gentagelser er de relative hyppigheder tættere samlet (ligger mellem 48,6% og 50,9%) end hvis man kun kastede 100 gange (mellem 40% og 55%). Hvis terningen var en

ægte terning(fuldstændig homogen og symmetrisk), måtte man på forhånd forvente, at det tal, som de relative hyppigheder grupperede sig omkring, var tallet 0.5. Man vil derfor sige, at sandsynligheden for at få et lige øjental er 0.5, eller kort $P(A) = 0.5$. ◆

1.3 Grundlæggende begreber

I det følgende vil eksempel 1.2 blive benyttet til illustration af definitioner og begreber.

Eksempel 1.2. Gennemgående eksempel.

En fabrik har erfaring for, at den daglige produktion af glasfigurer indeholder 10 % misfarvede, 20% har ridser, og 1 % af produktionen er både ridsede og misfarvede.
Et eksperiment består i tilfældigt at udtage en glasfigur af produktionen. Lad A være hændelsen at få en misfarvet og lad B være hændelsen at få en ridset.

◆

Komplemtærmængden til A benævnes $\bar{A}$ eller $\bar{A}$ og er mængden af alle udfald i udfaldsrummet U, der **ikke** er i A (den skraverede mængde på figur 1.1).

Eksempelvis er $\bar{A}$ i eksempel 1.2 mængden af alle glasfigurer, der ikke er misfarvet.
Idet $P(A) = 0.1$ ses umiddelbart, at $P(\bar{A}) = 1 - P(A) = 0.9$.

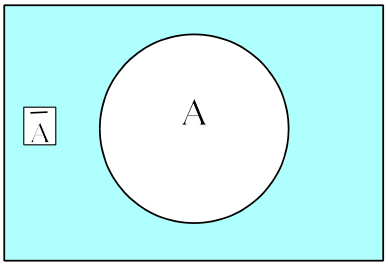


Fig. 1.1. *Komplementærmængde*

Vi har derfor klart følgende sætning: $\boxed{P(\bar{A}) = 1 - P(A)}$

Fællesmængden til A og B benævnes $A \cap B$ og er mængden af alle udfald i udfaldsrummet U, der tilhører **både A og B** (Den skraverede mængde i figur 1.2).

Eksempelvis er $A \cap B$ i eksempel 1.2 mængden af alle glasfigurer, der både er misfarvede og ridsede.
Af eksemplet følger, at $P(A \cap B) = 0.01$.

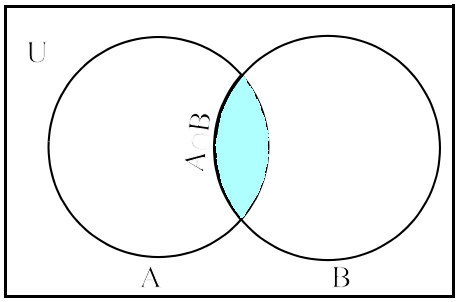


Fig 1.2. *Fællesmængde*

Foreningsmængden af A og B benævnes $A \cup B$ og er mængden af alle udfald i udfaldsrummet U, der **enten tilhører A eller B eventuelt dem begge** (den skraverede mængde på figur 1.3)

Eksempelvis er $A \cup B$ i eksempel 2.1 mængden af alle glasfigurer, der enten er misfarvede eller ridsede eventuelt begge dele.

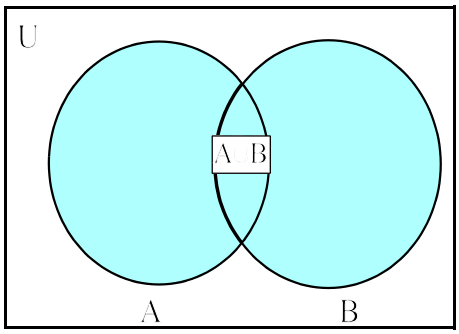


Fig. 1.3 *Foreningsmængde*

U kaldes for **den sikre hændelse**, dens komplementære hændelse for **den umulige hændelse** $\emptyset$.
$\boldsymbol{P(\emptyset) = 0}$
En hændelse, som kun indeholder ét udfald, kaldes en **elementarhændelse**.
To mængder hvis fællesmængde er tom kaldes **disjunkte.**

1.4. Additionssætning

Ved betragtning af fig.1.4 ses umiddelbart ved at betragte de skraverede arealer, at der gælder

Additionssætning: $P(A \cup B) = P(A) + P(B) - P(A \cap B).$

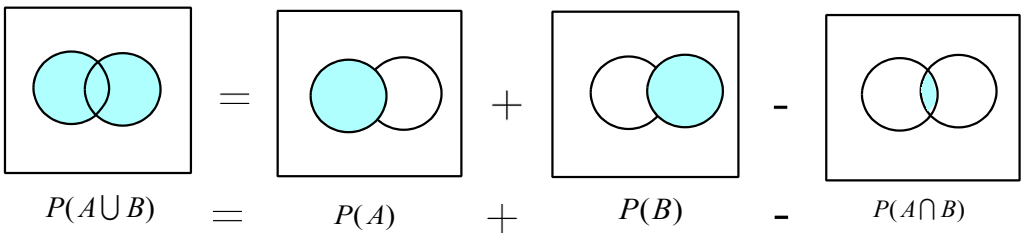


Fig.1.4 *Additionssætning*

I eksempel 1.2 er $P(A \cup B) = 0.1 + 0.2 - 0.01 = 0.29.$

1.5. Betinget sandsynlighed

For nogle hændelser A og B gælder, at $P(A \cap B) = P(A) \cdot P(B)$, men denne formel gælder ikke generelt. Eksempelvis er i eksempel 1.2 $P(A) \cdot P(B) = 0.1 \cdot 0.2 = 0.02 \neq P(A \cap B)$.

For at få en mere generel regel indføres $P(B \mid A)$ som kaldes sandsynligheden for, at B indtræffer, når A er indtruffet (den af A **betingede sandsynlighed** for B).

For at forklare den følgende definition, vil vi simplificere eksempel 1.2, idet vi antager, at den daglige produktion er 100 glasfigurer. I så fald er der 10 misfarvede figurer, 20 ridsede figurer, og 1 figur der er både misfarvet og ridset.
Hvis vi begrænser vort udfaldsrum til A, så er

$$P(B \mid A) = \frac{1}{10} = \frac{\frac{1}{100}}{\frac{10}{100}} = \frac{P(A \cap B)}{P(A)}.$$

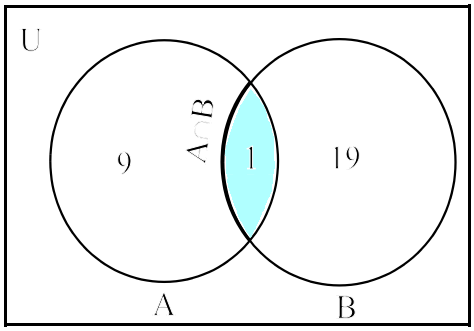


Fig. 1.5 ***Taleksempel***

Denne beregning begrunder rimeligheden i følgende definition:

Den af A **betingede sandsynlighed** for B $P(B \mid A)$ (eller sandsynligheden for, at B indtræffer, når A er indtruffet) defineres ved $P(B \mid A) = \dfrac{P(A \cap B)}{P(A)}$.

Ved multiplikation fås

Produktsætningen: $P(A \cap B) = P(A) \cdot P(B \mid A)$.

Benyttes produktsætningen på eksempel 1.2 fås $P(A \cap B) = P(A) \cdot P(B \mid A) = 0.1 \cdot 0.1 = 0.01$.

Eksempel 1.3: Betinget sandsynlighed.
En beholder indeholder 3 røde og 3 hvide kugler. Vi udtrækker successivt 2 kugler fra urnen.
Vi betragter følgende 2 hændelser:
A: Den først udtrukne kugle er rød. B: Den anden udtrukne kugle er rød.
Beregn $P(A \cap B)$ hvis
1) kugleudtrækningen foregår, ved at den først udtrukne kugle lægges tilbage før den anden udtrækkes.
2) kugleudtrækningen foregår, ved at den først udtrukne kugle **ikke** lægges tilbage før den anden udtrækkes.

Løsning
1) Her er $P(B \mid A) = \frac{3}{6}$ og derfor ifølge produktsætningen $P(A \cap B) = P(A) \cdot P(B \mid A) = \frac{1}{4}$
2) Her er $P(B \mid A) = \frac{2}{5}$ og derfor $P(A \cap B) = \frac{3}{6} \cdot \frac{2}{5} = \frac{1}{5}$

Bayes sætning

For to hændelser A og B for hvilken $P(A) > 0$ gælder

Bayes sætning: $P(B|A) = \dfrac{P(B) \cdot P(A|B)}{P(A)}$

Bevis:

Af definitionen på betinget sandsynlighed og produktsætningen fås $P(B|A) = \dfrac{P(A \cap B)}{P(A)} = \dfrac{P(B \cap A)}{P(A)} = \dfrac{P(B) \cdot P(A|B)}{P(A)}$

♦

Bayes sætning gør, at det er let at omskrive fra den ene betingende sandsynlighed til den anden. Dette er tilfældet, hvis den ene af de to betingede sandsynligheder $P(B|A)$ og $P(A|B)$ er meget lettere at beregne end den anden.

Eksempel 1.4 (Bayes sætning)

I en officeruddannelse kan man vælge mellem en "teknisk" linie og en "operativ"linie. På en bestemt årgang har 60 % valgt den operative linie og af disse er 20% kvinder. På den tekniske linie er 10% kvinder.

Ved lodtrækning vælges en elev.

a) Find sandsynligheden for, at denne er en kvinde.

Ved ovenstående lodtrækning viste det sig at eleven var en kvinde.

b) Hvad er sandsynligheden for, at hun kommer fra den tekniske linie.

Løsning:

Vi definerer følgende hændelser:

T: Den udtrukne er tekniker

K: Den udtrukne er en kvinde.

a) $P(K) = P(T \cap K) + P(O \cap K) = P(K|T) \cdot P(T) + P(K|O) \cdot P(O) = 0.1 \cdot 0.4 + 0.2 \cdot 0.6 = \underline{\underline{0.16 = 16\%}}$

b) Af Bayes sætning fås: $P(T|K) = \dfrac{P(K|T) \cdot P(T)}{P(K)} = \dfrac{0.1 \cdot 0.4}{0.16} = \underline{\underline{\dfrac{1}{4} = 25\%}}$

En anden metode ville det være, at antage, at der bliver optaget 100 elever.

Vi har så følgende skema

	Kvinder	I alt
Operativ	12	60
Teknisk	4	40

Heraf fås umiddelbart $P(K) = \dfrac{16}{100} = 16\%$ og $\underline{\underline{P(T|K) = \dfrac{4}{16} = 25\%}}$

♦

1.6. Statistisk uafhængighed.

To hændelser A og B siges at være statistisk uafhængige, såfremt $P(A \cap B) = P(A) \cdot P(B)$. Navnet skyldes, at vi i dette tilfælde har $P(B \mid A) = P(B)$ og $P(A \mid B) = P(A)$, således at sandsynligheden for, at den ene hændelse indtræffer, ikke afhænger af, om den anden hændelse indtræffer.

Eksempelvis ved kast med en terning, så vil sandsynligheden for at få en sekser i andet kast være uafhængigt af udfaldet i første kast, således at sandsynligheden for at få 2 seksere i de første 2 kast er $\frac{1}{6} \cdot \frac{1}{6}$.

Definitionen af statistisk uafhængighed generaliseres til flere hændelser end 2, således at der i tilfælde af 3 uafhængige hændelser A_1, A_2 og A_3 også gælder:
$P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2) \cdot P(A_3)$.

Opgaver

Opgave 1.1

I en mindre by viser en undersøgelse, at 60% af alle husstande holder en lokal avis, mens 30% holder en landsdækkende avis. Endvidere holder 10% af husstandene begge aviser.
Lad en husstand være tilfældig udvalgt, og lad A være den hændelse, at husstanden holder en lokal avis, og B den hændelse, at husstanden holder en landsdækkende avis.

Beregn sandsynlighederne for nedenstående hændelser.

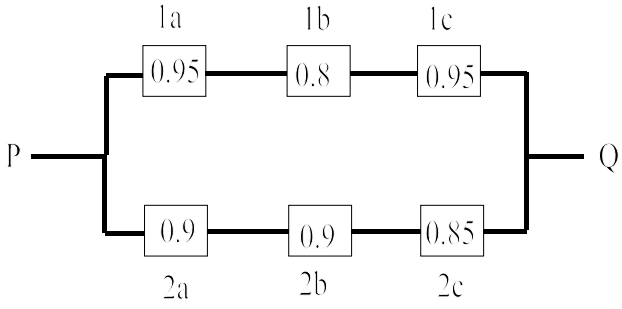
C: Husstanden holder begge aviser .
D: Husstanden holder kun den lokale avis.
E: Husstanden holder mindst én af aviserne.
F: Husstanden holder ingen avis
G: Husstanden holder netop én avis.

Det vides nu, at den tilfældigt valgte husstand holder den landsdækkende avis.

H: Husstanden holder den lokale avis.

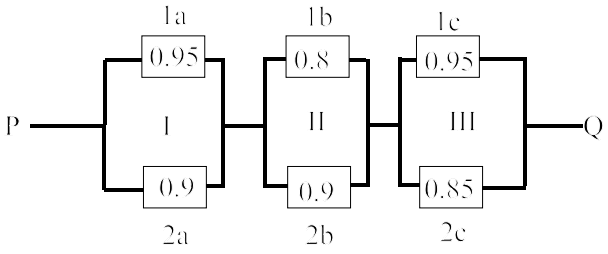
Opgave 1.2

1) I figur 1 er vist et elektrisk apparat, som kun fungerer, hvis enten alle komponenter 1a, 1b og 1c i den øverste ledning eller alle komponenter 2a, 2b og 2c i den nederste ledning fungerer. Sandsynligheden for at hver komponent fungerer er vist på tegningen, og det antages, at sandsynligheden for at en komponent fungerer er uafhængig af om de øvrige komponenter fungerer.



Figur 1

1) Hvad er sandsynligheden for at apparatet i figur 1 fungerer.



Figur 2

2) I figur 2 er vist et andet elektrisk apparat, som tilsvarende kun fungerer, hvis alle de tre kredsløb I, II og III fungerer, og det er kun tilfældet hvis enten den øverste eller den nederste komponent fungerer.
Hvad er sandsynligheden for at apparatet i figur 2 fungerer.

Opgave 1.3
Tre skytter skyder hver ét skud mod en skydeskive. De har træffesandsynligheder 0.75, 0.50 og 0.30.
Beregn sandsynligheden for
1) ingen træffere, 2) én træffer, 3) to træffere, 4) tre træffere.

Opgave 1.4
En "terning" har form som et regulært polyeder med 20 sideflader. På 4 sideflader er der skrevet 1, på 8 sideflader er der skrevet 6 mens der er skrevet 2, 3 , 4 og 5 på hver 2 sideflader.
Find sandsynligheden for i tre kast med denne terning at få
1) tre seksere
2) mindst én sekser
3) enten tre seksere eller tre enere

Opgave 1.5
En virksomhed fremstiller en bestemt slags apparater. Hvert apparat er sammensat af 5 komponenter. Heraf er 3 tilfældigt udvalgt blandt komponenter af typen a og 2 blandt komponenter af typen b. Det vides, at 10% af a-komponenterne er defekte og 20% af b-komponenterne er defekte. Et apparat fungerer hvis og kun hvis det ikke indeholder nogen defekt komponent.
Der udtages på tilfældig måde et apparat fra produktionen. Lad os betragte hændelserne:
A: Det udtagne apparat indeholder mindst 1 defekt a-komponent.
B: Det udtagne apparat indeholder mindst 1 defekt b-komponent.
1) Find $P(A)$, $P(B)$ og $P(A \cap B)$.
2) Find sandsynligheden for, at et apparat, der på tilfældig måde udtages af produktionen ikke fungerer.
3) Et apparat udtages på tilfældig måde fra produktionen og det konstateres ved afprøvning at det ikke fungerer. Find sandsynligheden for, at apparatet ikke indeholder nogen defekt a-komponent.

Opgave 1.6
To skytter konkurrerer ved en turnering. De har hver én patron og skyder mod en skive som giver 10 point, hvis et centralt område af skiven rammes og ellers 5 point. Rammes skiven ikke noteres 0 point.
Skytte A's dygtighed kan beskrives ved, at han i et skud har samme sandsynlighed for at få 10 points, 5 points eller 0 points.
Skytte B er dygtigere, idet hans sandsynligheder for at ramme er givet ved

Points y	10	5	0
$P(y)$	0.6	0.3	0.1

B har imidlertid fået en defekt patron med, der har sandsynligheden 50% for at fungere.
1) Beregn sandsynligheden for, at A vinder.
2) Det oplyses, at A vandt konkurrencen. Beregn sandsynligheden for, at B opnåede 5 points.

2. Kombinatorik

2.1. Indledning:

Såfremt et udfaldsrum U indeholder n udfald som alle er lige sandsynlige, vil sandsynligheden for hvert udfald være $P(u) = \frac{1}{n}$.

En hændelse A som indeholder a udfald vil da have sandsynligheden $P(A) = \frac{a}{n}$.

Dette udtrykkes ofte kort ved at sige, at sandsynligheden for A er antal gunstige udfald i A divideret med det totale antal udfald i udfaldsrummet.

I sådanne tilfælde, bliver problemet derfor, hvorledes man let kan optælle antal udfald. Dette kan ofte gøres ved benyttelse af **kombinatorik**.

2.2. Multiplikationsprincippet

Multiplikationsprincippet: Lad et valg bestå af n delvalg, hvoraf det første valg har r_1 valgmuligheder, det næste valg har r_2 valgmuligheder, . . . og det n'te valg har r_n valgmuligheder.

Det samlede antal valgmuligheder er da $r_1 \cdot r_2 \cdot \ldots \cdot r_n$

Multiplikationsprincippet illustreres ved følgende eksempel.

Eksempel 2.1

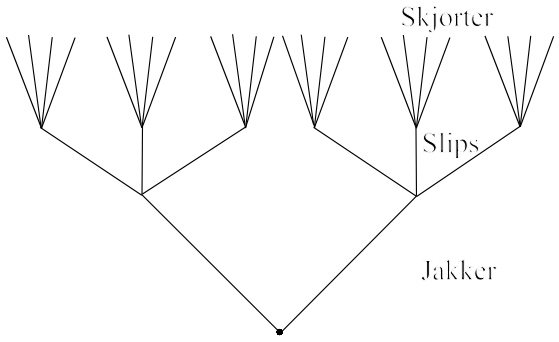
En mand ejer 2 forskellige jakker, 3 slips og 4 forskellige fabrikater skjorter.
På hvor mange forskellige måder kan han sammensætte sin påklædning af jakke, slips og skjorte.

Løsning:

1) Valg af jakke giver 2 valgmuligheder
2) Valg af slips giver 3 valgmuligheder
3) Valg af skjorte giver 4 valgmuligheder

Ifølge multiplikationsprincippet giver det i alt $2 \cdot 3 \cdot 4 = \underline{\underline{24}}$ muligheder

Man kunne illustrere løsningen ved følgende "forgreningsgraf"



2.3 Ordnet stikprøveudtagelse

Lad os tænke os vi har en beholder indeholdende 9 kugler med numrene 1, 2, 3, ..., 9 .
Vi udtager nu en stikprøve på 4 kugler. Det kan ske

1) uden tilbagelægning: En kugle er taget op, nummeret noteres, men den lægges ikke tilbage inden man tager en ny kugle op.
2) med tilbagelægning: En kugle tages op, nummeret noteres, og derefter lægges kuglen tilbage inden man tager en ny kugle op. Man kan følgelig få den samme kugle op flere gange.

Ved en ordnet stikprøveudtagelse lægges vægt på den rækkefølge hvori kuglerne udtages, .
dvs. der er forskel på 2,1,3,5 og 3,1,2,5

2.3.1 Uden tilbagelægning

Eksempel 2.2. Ordnet uden tilbagelægning

I en forening skal der blandt 10 kandidater vælges en bestyrelse
På hvor mange forskellige måder kan man sammensætte denne bestyrelse, hvis

1) Bestyrelsen består af en formand og en kasserer
2 Bestyrelsen består af en formand, en næstformand, en kasserer og en sekretær.

Løsning:

1) En formand vælges blandt 10 kandidater — 10 valgmuligheder
   En Kasserer vælges blandt de resterende 9 kandidater — 9 valgmuligheder
   Da der for hvert valg af formand er 9 muligheder for kasserer, følger af multiplikationsprincippet, at det totale antal forskellige bestyrelser er $10 \cdot 9 = \underline{\underline{90}}$.
2) Analogt fås ifølge multiplikationsprincippet at antal forskellige bestyrelser
   $10 \cdot 9 \cdot 8 \cdot 7 = \underline{\underline{5040}}$
   Ti 89: MATH \ Probability\ nPr \ ENTER . Resultat: nPr(10,4) = 5040

♦

Eksempel 2.2 begrunder følgende definition

Permutationer. Antal måder (rækkefølger eller "permutationer") som *m* elementer kan udtages (ordnet og uden tilbagelægning) ud af n elementer er $P(n,m) = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-m+1)$

***n* fakultet (*n* udråbstegn)** $n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1$
Endvidere defineres $0! = 1$.

Eksempel 2.3. *n* fakultet

En pentapeptide består af en kæde af følgende 5 aminosyrer: alanine, valine, glycine, cysteine, trypophan. Den har forskellige egenskaber afhængig af den rækkefølge de 5 aminosyrer sidder i kæden. Hvor mange forskellige typer pentapeptider er der mulig at danne ?

Løsning:

Da rækkefølgen spiller en rolle er antallet af pentapeptider = $\underline{\underline{5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}}$.

TI 89: 5 MATH \ Probability\ ! \ ENTER . Resultat: 5! = 120

♦

2.3.2 Med tilbagelægning

Eksempel 2.4. Ordnet, med tilbagelægning
I en forening skal 4 tillidshverv fordeles mellem 10 personer. En person kan godt have flere tillidshverv. På hvor mange forskellige måder kan disse hverv fordeles.?
Løsning:

Tillidshverv 1 placeres.	10 valgmuligheder
Tillidshverv 2 placeres	10 valgmuligheder
Tillidshverv 3 placeres	10 valgmuligheder
Tillidshverv 4 placeres	10 valgmuligheder
I alt (ifølge multiplicationsprincippet)	$10 \cdot 10 \cdot 10 \cdot 10 = 10^4$

♦

2.4. Uordnet stikprøveudtagelse

Eksempel 2.5 Uordnet uden tilbagelægning

En beholder indeholdende 5 kugler med numrene k_1, k_2, k_3, k_4, k_5

Vi udtager nu en stikprøve på 3 kugler uden tilbagelægning. Rækkefølgen kuglen tages op er uden betydning, dvs. der er ikke forskel på eksempelvis k_1, k_4, k_2 og k_4, k_1, k_2

Hvor mange forskellige stikprøver kan forekomme?
Løsning:
Antallet er ikke flere end man kan foretage en simpel optælling:
$\{k_1,k_2,k_3\},\{k_1,k_2,k_4\}\{k_1,k_2,k_5\}\{k_1,k_3,k_4\}\{k_1,k_3,k_5\}\{k_2,k_3,k_4\}\{k_2,k_3,k_5\}\{k_2,k_4,k_5\}\{k3,k_4,k_5\}$
Antal stikprøver = 10

♦

Det er klart, at ren optælling er uoverkommeligt, hvis mængden er stor.

Definition af kombination
Lad M være en mængde med *n* elementer.
En delmængde af M med *r* elementer kaldes en **kombination** af med *r* elementer fra M.
Antallet af kombinationer med *r* elementer betegnes K(*n*,*r*) eller $\binom{n}{r}$ (*n* over *r*).

Sætning 2.1 (Antal kombinationer).
Antal kombinationer med r elementer fra en mængde på n elementer er $K(n,r) = \frac{n!}{r! \cdot (n-r)!}$

Bevis: Beviset knyttes for enkelheds skyld til et taleksempel, som let kan generaliseres.
Lad os antage, vi på tilfældig måde udtager 3 kugler af en kasse, der indeholder 5 kugler med numrene k_1, k_2, k_3, k_4, k_5.

Vi skal nu vise, at $k(5,3) = \frac{5!}{3!\cdot 2!}$

Lad os først gå ud fra, at rækkefølgen hvori kuglerne trækkes er af betydning, Der er altså eksempelvis forskel på k_1, k_3, k_4 og k_3, k_1, k_4. Dette kan gøres på P(5,3) = $5 \cdot 4 \cdot 3$ måder.
Hvis de 3 kugler udtages, så rækkefølgen **ikke** spiller en rolle, har vi vedtaget, det kan gøres på K(5,3) måder. Lad en af disse måder være k_1, k_3, k_4. Disse 3 elementer kan ordnes i rækkefølge på $3! = 3 \cdot 2 \cdot 1$ måder.
Vi har følgelig, at $P(5,3) = K(5,3) \cdot 3! \Leftrightarrow K(5,3) = \frac{P(5,3)}{3!} \Leftrightarrow K(5,3) = \frac{5 \cdot 4 \cdot 3}{3!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{3! \cdot 2!} = \frac{5!}{3! \cdot 2!}$ ♦

Eksempel 2.6. Antal kombinationer

I en forening skal der blandt 10 kandidater vælges 4 personer til en bestyrelse
På hvor mange forskellige måder kan man sammensætte denne bestyrelse?

Løsning:

Antal måder man kan sammensætte bestyrelsen er

$$K(10,4) = \frac{10!}{4! \cdot 6!} = \frac{10 \cdot 9 \cdot 8 \cdot 7}{4!} = 10 \cdot 3 \cdot 7 = \underline{\underline{210}} \text{ måder}$$

TI 89: 5 MATH \ Probability\ nCr \ ENTER . Resultat: nCr(10,4) = 120

♦

Som tidligere nævnt finder man ofte sandsynligheden for en hændelse A ved at beregne antallet af gunstige udfald i A og dividere det med det totale antal udfald i udfaldsrummet.
Det følgende eksempel illustrerer dette.

Eksempel 2.7. Beregning af sandsynlighed

Fra et almindeligt spil kort (52 kort) trækkes 5 kort.
Find sandsynligheden for, at få pokermeldingen "Full-house" dvs. tre kort har samme talværdi og de to andre har samme talværdi (f.eks. tre femmere og to knægte).

Løsning:

Vælge 3 esser ud af 4 esser K(4,3) = 4 valgmuligheder
Da der er 13 kort i en farve, så kan man vælge 3 med samme talværdi på $13 \cdot 4 = 52$ måder.
Vælge 2 konger ud af 4 konger K(4,2) = 6 valgmuligheder
Da man ikke må vælge den talværdi, hvor man har udtrukket 3 kort fra,
så kan man vælge 2 med samme talværdi på $12 \cdot 6 = 72$ måder
Antal gunstige er derfor ifølge multiplikationsprincippet $52 \cdot 72 = 3744$
Det totale antal måder man kan udtage 5 kort ud af 52 kort er K(52,5) = 2598960

$$P(\text{Full House}) = \frac{3744}{2598960} = 0.00144 = \underline{\underline{0.144\%}}$$

2.5 Hypergeometrisk fordeling

Af særlig interesse er den såkaldte "hypergeometriske fordeling", som bl.a. finder anvendelse ved kvalitetskontrol af varepartier (jævnfør eksempel 2.9),ved markedsundersøgelser, hvor man uden tilbagelægning udtager en repræsentativ stikprøve på eksempelvis 500 personer

I det følgende eksempel "udledes" formlen for den hypergeometriske fordeling.

Eksempel 2.8. Hypergeometrisk fordeling

I en forening skal der blandt 5 kvindelige og 8 mandlige kandidater vælges en bestyrelse på 4 personer. Find sandsynligheden for, at der er netop 1 kvinde i bestyrelsen..

Løsning:

X = antal kvinder i bestyrelsen

At der skal være netop 1 kvinde i bestyrelsen forudsætter, at vi udtager 1 kvinde ud af de 5 kvinder og 3 mænd ud af de 8 mænd.

At udtage 1 kvinde ud af 5 kvinder kan gøres på K(5,1) måder

At udtage 3 mænd ud af 8 mænd kan gøres på K(8,3) måder.

Antal gunstige udfald er ifølge multiplikationsprincippet $K(5,1)\cdot K(8,3)$

Det totale antal udfald fås ved at udtage 4 personer ud af de 13 kandidater

Dette kan gøres på K(13,4) måder.

$$P(X=1)=\frac{K(5,1)\cdot K(8,3)}{K(13,4)}=\underline{\underline{0.3916}}$$

◆

Definition af hypergeometrisk fordeling.

Lad der i en beholder befinde sig N kugler, hvoraf M er defekte.

En kugle udtrækkes og undersøges.

Dette gentages n gange **uden mellemliggende tilbagelægning**

Lad X være antallet af defekte kugler, som udtrækkes. Der gælder da

$$P(X=x)=\frac{K(M,x)\cdot K(N-M,n-x)}{K(N,n)},\ x\in\{0,1,2,3,\ldots,M\}\cap\{0,1,2,3,\ldots,n\}$$

Eksempel 2.8 (fortsat)

I eksempel 2.8 benyttede vi den hypergeometriske fordeling i det tilfælde, hvor N = 13, n =4 , M = 5 og $x=1$. ◆

Eksempel 2.9. Kvalitetskontrol

En producent fabrikerer komponenter, som sælges i æsker med 600 komponenter i hver. Som led i en kvalitetskontrol udtages hvert kvarter tilfældigt en æske produceret indenfor de sidste 15 minutter, og 25 tilfældigt udvalgte komponenter i denne undersøges, hvorefter det foregående kvarters produktion godkendes, såfremt der højst er én defekt komponent i stikprøven.
Hvor stor er acceptsandsynligheden p, hvis æsken indeholder i alt 10 defekte komponenter, og udtrækningen sker **uden** mellemliggende tilbagelægning ?

Løsning:
Lad X være antallet af defekte blandt de 25 komponenter
Vi har: $p = P(X = 0) + P(X = 1)$.

$$P(X=0) = \frac{K(10,0)\cdot K(590,25)}{K(600,25)} = 0.6512 \text{ .og } P(X=1) = \frac{K(10,1)\cdot K(590,24)}{K(600,25)} = 0.2876 \text{ .}$$

Vi har altså p = 0.6512 + 0.2876 = 0.9388 = 93.88%. ◆

Opgaver

Opgave 2.1.

a) Bestem det antal måder, hvorpå bogstaverne A, B og C kan stilles rækkefølge.
b) Samme opgave for A, B, C og D.

Opgave 2.2.

På et spisekort er opført 6 forretter, 10 hovedretter og 4 desserter.

1) Hvor mange forskellige middage bestående enten af forret og hovedret eller af hovedret og dessert kan man sammensætte.
2) Hvor mange forskellige middage bestående af en forret, en hovedret og en dessert kan man sammensætte.

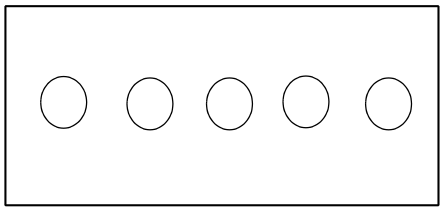
Opgave 2.3.

En test består af 40 spørgsmål, der alle skal besvares med ,'ja'. 'nej' og 'ved ikke'. På hvor mange forskellige måder kan prøven besvares?

Opgave 2.4.

I en virksomhed skal der installeres et kaldesystem. I hvert lokale opsættes et batteri af n lamper, og hver af de ansatte har sin bestemte lampekombination.

1) Hvis $n = 5$, hvor mange ansatte kan da have deres eget kaldesystem (se figuren)
2) Hvis virksomheden har 500 ansatte, hvor stor skal n så være.



Opgave 2.5

Normale personbilers indregistreringsnumre består af to bogstaver og et nummer mellem 20000 og 59999 .
Lad os antage, at man er nået til numre der begynder med UV. Et eksempel på en nummerplade er da UV 54755
Hvad er sandsynligheden for, at en nyindregistreret bil får et registreringsnummer med lutter forskellige cifre, når vi antager, at alle cifre har samme sandsynlighed?

Opgave 2.6

Til en julemiddag er der dækket til familiens 12 medlemmer ved et langt bord.

1) På hvor mange måder kan de placere sig ved bordet?

Efter middagen danner hele familien kæde og danser omkring juletræet.

2) På hvor mange måder kunne de danne kæde?

Opgave 2.7

En klasse med 21 elever skal under en øvelse fordeles på 5 grupper. 4 af grupperne skal være på 4 elever, og 1 gruppe skal være på 5 elever.
På hvor mange måder kan fordelingen af eleverne på de 5 grupper foregå?

Opgave 2.8
Af en forsamling på 8 kvinder og 12 mænd skal udtages et udvalg på 5 medlemmer.
På hvor mange måder kan dette ske, når udvalget skal indeholde
1) Mindst et medlem af hvert sit køn
2) Mindst 2 kvinder og mindst 2 mænd.

Opgave 2.9.
Bestem antallet af 5-cifrede tal, der kan skrives med to l-taller, et 2- tal og to 3-taller.

Opgave 2.10
En beholder indeholder 3 hvide, 6 røde og 3 sorte kugler
3 kugler udtrækkes tilfældigt uden tilbagelægning.
Find sandsynligheden for at de er af samme farve.

Opgave 2.11
Fra et sædvanligt spil kort udtrækkes på tilfældig måde 3 kort uden tilbagelægning. Bestem sandsynlighederne for hver af hændelserne
A: Der udtrækkes kun 8'ere.
B: Der udtrækkes lutter hjerter.
C: Der udtrækkes 2 sorte og 1 rødt kort.

Opgave 2.12
Ved en lodtrækning fordeles 3 gevinster blandt 25 lodsedler. En spiller har købt 5 lodsedler. Beregn sandsynligheden for hver af følgende hændelser:
1) Spilleren vinder alle tre gevinster.
2) Spilleren vinder ingen gevinster.
3) Spilleren vinder netop én gevinst.

Opgave 2.13
I en urne findes 2 blå, 3 røde og 5 hvide kugler. 3 gange efter hinanden optages tilfældigt en kugle fra urnen uden mellemliggende tilbagelægning.
1) Find sandsynligheden for hændelsen A, at der højst optages 2 hvide kugler,
2) Find sandsynligheden for hændelsen B, at de optagne kugler har hver sin farve.
3) Find sandsynligheden for, at de tre kugler har samme farve,

Opgave 2.14
En fabrikant fremstiller en bestemt type radiokomponenter. Disse leveres i æsker med 30 komponenter i hver æske. En køber har den aftale med fabrikanten, at hvis en æske indeholder 4 defekte komponenter eller derover, kan køberen returnere æsken, i modsat fald skal den godkendes. Køberen kontrollere hver æske ved en stikprøve, idet han af æsken udtager 10 komponenter tilfældigt. Lad X være antal defekte i stikprøven. Der overvejes nu to planer:
1) Hvis $X = 0$, så godkendes æsken, ellers undersøges æsken nærmere.
2) Hvis $X \leq 1$, så godkendes æsken, ellers undersøges æsken nærmere.
Hvad er sandsynligheden for, at en æske, der indeholder netop 4 defekte komponenter, bliver godkendt af køberen ved metode 1 og ved metode 2.

3 Stokastisk variabel

3.1. Definition af stokastisk variabel

Skal man behandle et problem statistisk må det på en eller anden måde være muligt at behandle det talmæssigt.
Betragtes således et eksempel med kast med en mønt, kunne man til udfaldet plat tilordne tallet 0 og til udfaldet krone tilordne tallet 1 og på den måde få problemet overført til noget, hvor man kan foretage beregninger. Man siger, man har indført en stokastisk (eller statistisk) variabel X, som er 0, når udfaldet er plat, og 1 når udfaldet er krone.
Generelt gælder følgende definition:

DEFINITION af stokastisk variabel (engelsk: random variable).
En stokastisk variabel X er en funktion, som til hvert udfald i udfaldsrummet lader svare et reelt tal.

En stokastisk variabel betegnes med et stort bogstav såsom X, mens det tilsvarende lille bogstav x betegner en mulig værdi af X.

Udtager man således en stikprøve på 10 møtrikker ud af en kasse på 100 møtrikker, kunne X eksempelvis være defineret som " antal defekte møtrikker blandt de 10".
Eksperimenterer man med at anvende en ny metode til fremstilling af et produkt, kunne X være "det målte procentiske udbytte ved forsøget".

Ved en **diskret** variabel forstås en variabel, hvis mulige værdier udgør en endelig eller tællelig mængde. I eksemplet hvor X er antal defekte møtrikker, er X en diskret variabel, da den kun kan antage heltallige værdier fra 0 til 10. Diskrete variable kaldes ofte for tællevariable, da de ofte optræder ved optællinger, for eksempel som ovenfor i forbindelse med kvalitetskontrol.

Ved en **kontinuert** variabel forstås en variabel, hvis mulige værdier er alle reelle tal i et vist interval. I eksemplet, hvor Y er det målte procentiske udbytte, er Y en kontinuert variabel, da den kan antage alle værdier fra 0% til 100%.

3.2 Sandsynlighedsfordeling for diskret stokastisk variabel

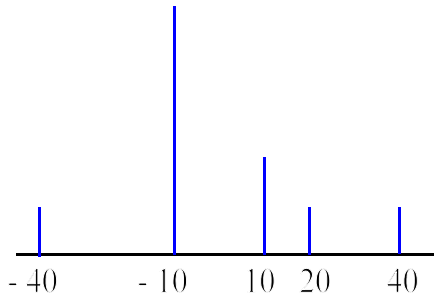
Vi vil i dette afsnit benytte følgende eksempel 3.1 til illustration af definitioner og begreber.

Eksempel 3.1. Diskret variabel.
Ved et roulettespil vil resultatet "rød" give en gevinst på 40 kr. Tilsvarende vil "grøn" og "blå" give gevinster på henholdsvis 20 og 10 kr, mens "gul" og "sort" giver tab på henholdsvis 10 kr og 40 kr.
Lad X være gevinsten ved et spil. X får da værdierne 40, 20, 10, -10 og -40.
Ejeren af spillet ved, at $P(X = 40) = 0.1$, $P(X = 20) = 0.1$, $P(X = 10) = 0.2$, $P(X = -10) = 0.5$ og $P(X = -40) = 0.1$.

Man siger, at den diskrete variabel X har en **sandsynlighedsfordeling eller en "tæthedsfunktion"** givet ved ovenstående.
Sædvanligvis angives sandsynlighedsfordelingen ved en tabel:

u	sort	gul	blå	grøn	rød
x	-40	-10	10	20	40
$P(X=x)$	0.1	0.5	0.2	0.1	0.1

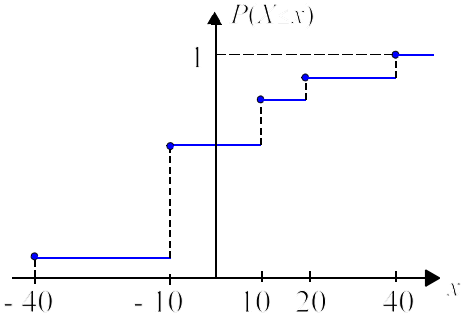
En sandsynlighedsfordeling illustreres grafisk ved at tegne et stolpediagram som vist på figuren.



Stolpediagram for tæthedsfunktion

I visse situationer er det en fordel at betragte den "kumulerede" sandsynlighedsfunktion for X, der kaldes X's **fordelingsfunktion** $F(x) = P(X \le x)$.
En tabel for denne er

u	sort	gul	blå	grøn	rød
x	-40	-10	10	20	40
$P(X=x)$	0.1	0.5	0.2	0.1	0.1
$P(X \le x)$	0.1	0.6	0.8	0.9	1.00



Trappekurve for fordelingsfunktion

3.3. Middelværdi

Ejeren ønsker at finde ud af, hvor stor en gevinst pr. spil han i middel kan forvente. Hertil må "middelværdien" beregnes.

Eksempel 3.2 Anskuelig beregning af middelværdi
Hvis vi ved roulettespillet i eksempel 3.1 tænker os at vi har spillet 100 spil, ville vi med de givne sandsynligheder forvente, at resultatet i middel bliver følgende
$0.1 \cdot 100 = 10$ gange tabes 40 kr dvs. ialt vindes $-0.1 \cdot 100 \cdot 40 = -400$ kr
$0.5 \cdot 100 = 50$ gange tabes 10 kr dvs. i alt vindes $-0.5 \cdot 100 \cdot 10 = -500$ kr
$0.2 \cdot 100 = 20$ gange vindes 10 kr dvs. ialt vindes $0.2 \cdot 100 \cdot 10 = 200$ kr
$0.1 \cdot 100 = 10$ gange vindes 20 kr dvs. ialt vindes $0.1 \cdot 100 \cdot 20 = 200$ kr
$0.1 \cdot 100 = 10$ gange vindes 40 kr dvs. i alt vindes $0.1 \cdot 100 \cdot 40 = 400$ kr
Vi ville derfor på 100 spil i alt vinde -100 kr eller altså i middel tabe 1 kr, eller ejeren vil i middel vinde 1 kr pr. spil.
Beregningen kunne også skrives $\frac{-0.1 \cdot 100 \cdot 40 - 0.5 \cdot 100 \cdot 10 + 02 \cdot 100 \cdot 10 + 0.1 \cdot 100 \cdot 20 + 0.1 \cdot 100 \cdot 40}{100}$

$= -0.1 \cdot 40 - 0.5 \cdot 10 + 02 \cdot 10 + 0.1 \cdot 20 + 0.1 \cdot 40 = -1$ eller
$-40 \cdot P(X = -40) + (-10) \cdot P(X = -10) + 10 \cdot P(X = 10) + 20 \cdot P(X = 20) + 40 \cdot P(X = 40)$

◆

Eksempel 3.2 begrunder følgende definition af middelværdi for en diskret variabel X:

DEFINITION af middelværdi for diskret variabel.
Middelværdi for en diskret variabel X benævnes μ *eller* $E(X)$ *og er defineret som*

$$\mu = E(X) = x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) + \ldots + x_n \cdot P(X = x_n)$$

eller kort $\mu = E(X) = \sum_{i=1}^{n} x_i \cdot P(X = x_i)$ (*E står for expected value*)

Regneregler for middelværdi.
Lad X og Y være stokastiske variable knyttet til samme udfaldsrum U og lad a og b være konstanter.
Der gælder da $E(aX + b) = a \cdot E(X) + b$ (1)
og $E(X + Y) = E(X) + E(Y)$ (2)

Bevis:
Vi vil bevise reglerne ved at gå ud fra et taleksempel, som så let kan generaliseres.
1) Bevis for (1)
Lad os antage, at ejeren af det i eksempel 3.2 nævnte roulettespil påtænker at ændre gevinsterne efter forskriften $Z = 5 \cdot X - 10$
Vi vil vise, at $E(Z) = E(5 \cdot X + b) = 5 \cdot E(X) - 10$

Vi får følgende tabel over sandsynlighedsfordelingen:

u	sort	gul	blå	grøn	rød
x	-40	-10	10	20	40
z	$5\cdot(-40)-10=-210$	$5\cdot(-10)-10=-60$	$5\cdot(10)-10=40$	$5\cdot 20-10=90$	$5\cdot 40-10=190$
$P(X=x)$ $=P(Z=z)$	0.1	0.5	0.2	0.1	0.1

$$E(Z)=(-210)\cdot 0.1+(-60)\cdot 0.5+40\cdot 0.2+90\cdot 0.1+190\cdot 0.1$$
$$=(5\cdot(-40)-10)\cdot 0.1+(5\cdot(-10)-10)\cdot 0.5+(5\cdot 10-10)\cdot 0.2+(5\cdot 20-10)\cdot 0.1+(5\cdot 40-10)\cdot 0.1$$
$$=5\cdot\left[(-40)\cdot 0.1+(-10)\cdot 0.5+10\cdot 0.2+20\cdot 0.1+40\cdot 0.1\right]+(-10)\cdot\left[0.1+0.5+0.2+0.1+0.1\right]=5\cdot E(X)-10$$

2) Bevis for (2)
Lad os antage, at ejeren af det i eksempel 3.2 nævnte roulettespil yderligere køber en roulette, men her ændrer gevinsterne y som det fremgår af følgende tabel

u	sort	gul	blå	grøn	rød
x	-40	-10	10	20	40
y	-50	0	15	30	45
$P(X=x)$ $=P(Y=y)$	0.1	0.5	0.2	0.1	0.1

$$E(X+Y)=(-40-50)\cdot 0.1+(-10-5)\cdot 0.5+(10+15)\cdot 0.2+(20+30)\cdot 0.1+(40+45)\cdot 0.1$$
$$=(-40)\cdot 0.1+(-10)\cdot 0.5+10\cdot 0.2+20\cdot 0.1+40\cdot 0.1+\left[(-50)\cdot 0.1+(-5)\cdot 0.5+15\cdot 0.2+30\cdot 0.1+45\cdot 0.1\right]=E(X)+E(Y)$$

3.4. Varians og spredning

Udover at angive en middelværdi, er man ofte også interesseret i at angive et mål for om værdierne ligger tæt omkring middelværdien, eller om de spreder sig meget (varierer meget). Som et mål herfor beregnes varians og spredning for variablen X.
Desværre har disse tal ikke samme umiddelbare fortolkning som middelværdien.
Det er dog klart, at hvis man i en opinionsundersøgelse spørger 1000 vælgere (repræsentativt udvalgt) om hvilket parti de vil stemme på hvis der var valg i morgen, og på den basis konkluderer at 30% vil stemme på partiet venstre, så er det væsentlig at vide hvor usikkert dette tal er, og i en sådan situation må man kende et spredningsmål.
Vi vil igen illustrere dette ved tallene fra eksempel 3.1

Eksempel 3.3 . Varians og spredning
I eksempel 3.1 fandt vi følgende sandsynlighedsfunktion

u	sort	gul	blå	grøn	rød
x	-40	-10	10	20	40
$P(X=x)$	0.1	0.5	0.2	0.1	0.1

og vi fandt, at middelværdien $\mu = -1$
Som et spredningsmål kunne man beregne middelværdien af de enkelte værdiers afvigelser fra middelværdien, men da den altid ville være 0 betragtes i stedet middelværdien af de kvadratiske afvigelser altså $E\left((X-\mu)^2\right)$
Dette tal kaldes variansen af X
$V(X) = (-40-(-1))^2 \cdot 0.1 + (-10-(-1))^2 \cdot 0.5 + (10-(-1))^2 \cdot 0.2 + (20-(-1))^2 \cdot 0.1 + (40-(-1))^2 \cdot 0.1 = 429$
Da variansens enhed jo er kr^2, og man sædvanligvis ønsker at måle spredningen i kr, så defineres spredningen $\sigma(X) = \sqrt{V(X)} = \sqrt{429} = 20.71$

◆

Som det fremgår af eksempel 3.3 gælder følgende definitioner:

DEFINITION af varians og spredning for diskret variabel. Variansen for en diskret variabel X benævnes σ^2 eller V (X) og er defineret som
$\sigma^2 = V(X) = E\left((X-\mu)^2\right) = \sum_x (x-\mu)^2 \cdot P(X=x)$.
Spredningen (engelsk: standard deviation) benævnes σ og er defineret som $\sigma = \sqrt{V(X)}$.

Alle x- værdier i eksempel 1.2 ligger indenfor en afstand af $2 \cdot \sigma$ fra middelværdien -1.
For de fleste fordelinger vil mindst 99% af alle værdier ligge indenfor $[\mu - 3\cdot\sigma\,;\,\mu + 3\cdot\sigma]$, og sædvanligvis vil mindst 95% af værdierne ligge indenfor $[\mu - 2\cdot\sigma\,;\,\mu + 2\cdot\sigma]$.

At anvende definitionen på varians er ofte regnemæssigt besværligt. Den følgende formel giver enklere regninger

$$V(X) = E\left(X^2\right) - \left(E(X)\right)^2$$

Bevis:

$V(X) = E\left(X-\mu)^2\right) = E\left(X^2 - 2\cdot\mu\cdot X + \mu^2\right)$ (Parentesen er udregnet)
$= E\left(X^2\right) + E(-2\mu X) + E\left(\mu^2\right)$ (regneregel for middelværdi nr. (2))
$= E\left(X^2\right) - 2\mu\cdot E(X) + E\left(\mu^2\right)$ (regneregel for middelværdier nr. (1))
$= E\left(X^2\right) - 2\mu\cdot\mu + \mu^2$ (regneregel for middelværdier nr. (1))
$= E\left(X^2\right) - \mu^2 = E\left(X^2\right) - \left(E(X)\right)^2$

◆

For fordelingen i eksempel 3.2 fås
$V(X) = (-40)^2 \cdot 0.1 + (-10)^2 \cdot 0.5 + 10^2 \cdot 0.2 + 20^2 \cdot 0.1 + 40^2 \cdot 0.1 - (-1)^2 = 429$

Regneregler for varians.
Lad X være en stokastisk variabel og lad a og b være konstanter.
Der gælder da $V(aX+b) = a^2 \cdot V(X)$

Bevis:

$V(aX+b) = E\left((aX+b)^2\right) - \left(E(aX+b)\right)^2$ (følger af formlen $V(X) = E\left(X^2\right) - \left(E(X)\right)^2$)

$= E\left(a^2X^2 + 2abX + b^2\right) - \left(aE(X)+b\right)^2$ (udregne $(aX+b)^2$ +regneregel for middelværdi nr (1))

$= a^2 \cdot E(X^2) + 2abE(X) + b^2 - \left(a^2\left(E(X)\right)^2 + 2abE(X) + b^2\right)$ (regneregel for middelværdi nr (1)

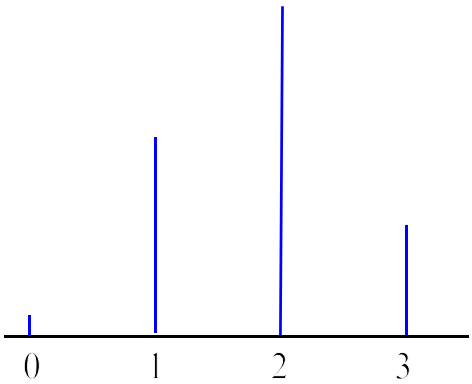
$= a^2 \cdot E(X^2) - a^2\left(E(X)\right)^2 = a^2 \cdot \left(E(X^2) - a^2\left(E(X)\right)^2\right) = a^2 \cdot V(X)$ ◆

Den hypergeometriske fordeling har vi behandlet i kapitel 2, men ikke der forklaret hvad der lå i ordet "fordeling".
Dette vil nu fremgå af det følgende eksempel.

Eksempel 3.4: Hypergeometrisk fordeling.
I en forening har 10 medlemmer ønsket at blive valgt til bestyrelsen. Af disse er de 6 mænd, og de 4 er kvinder. Bestyrelsen er på 3 medlemmer.
Lad X betegne antallet af mænd i bestyrelsen
Find og skitser tæthedsfunktionen for X, og beregn middelværdi og spredning for X.
Løsning:
X er en diskret stokastisk variabel, der er hypergeometrisk fordelt.

Sandsynlighedsfordelingen fremgår af følgende skema:

x	0	1	2	3
$P(X=x)$	$\frac{K(6,0)\cdot K(4,3)}{K(10,3)} = \frac{4}{120}$	$\frac{K(6,1)\cdot K(4,2)}{K(10,3)} = \frac{36}{120}$	$\frac{K(6,20)\cdot K(4,1)}{K(10,3)} = \frac{60}{120}$	$\frac{K(6,3)\cdot K(4,0)}{K(10,3)} = \frac{20}{120}$



Stolpediagram for h (10, 6, 3).

$$E(X) = 0 \cdot \frac{4}{120} + 1 \cdot \frac{36}{120} + 2 \cdot \frac{60}{120} + 3 \cdot \frac{20}{120} = \frac{216}{120} = \underline{\underline{1.8}}$$

$$V(X) = E(X^2) - \mu^2 = 0^2 \cdot \frac{4}{120} + 1^2 \cdot \frac{36}{120} + 2^2 \cdot \frac{60}{120} + 3^2 \cdot \frac{20}{120} - 1.8^2 = 0.56$$

$$\sigma(X) = \sqrt{0.56} = \underline{\underline{0.748}}$$

Det kan vises, at hvis $p = \frac{M}{N}$ er middelværdien for $E(X) = n \cdot p$ og spredningen $\sigma(X) = \sqrt{n \cdot p \cdot (1-p) \cdot \frac{N-n}{N-1}}$.

I ovenstående tilfælde er $p = \frac{M}{N} = \frac{6}{10}$, $E(X) = n \cdot p = 3 \cdot \frac{6}{10} = \underline{\underline{1.8}}$ og

$$\sigma(X) = \sqrt{n \cdot p \cdot (1-p) \cdot \frac{N-n}{N-1}} = \sqrt{3 \cdot \frac{6}{10} \cdot \left(1 - \frac{6}{10}\right) \cdot \frac{10-3}{10-1}} = \underline{\underline{0.748}}$$

Opgaver

Opgave 3.1
Givet følgende to funktioner:

x	-3	-2	-1	3	6	ellers
X	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{1}{5}$	0
Y	$\frac{1}{10}$	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	0

1) Kun en af de to variable X og Y kan opfattes som en stokastisk variabel med den i tabellen angivne sandsynlighedsfordeling
2) For den i spørgsmål 1 fundne sandsynlighedsfunktion Z, skal man finde den tilsvarende fordelingsfunktion, og tegne henholdsvis stolpediagram og sumkurve.
3) Beregn middelværdien $E(Z)$, variansen $V(Z)$ og spredningen $\sigma(Z)$.

Opgave 3.2
Givet følgende funktion:

x	-3	-2	0	1	ellers
$f(x)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{6}$	$\frac{1}{3}$	0

1) Idet $f(x)$ opfattes som en tæthedsfunktion for en statistisk variabel X skal man tegne det tilhørende stolpediagram.
2) Find den tilsvarende fordelingsfunktion $F(x)$ og tegn grafen.
3) Beregn middelværdien $E(X)$, variansen $V(X)$ og spredningen $\sigma(X)$.

Opgave 3.3
Givet følgende fordelingsfunktion $F(x)$

x	x < -5	$-5 \le x < -3$	$-3 \le x < -1$	$-1 \le x < 1$	$1 \le x < 3$	$3 \le x$
$F(x)$	0	$\frac{1}{12}$	$\frac{4}{12}$	$\frac{8}{12}$	$\frac{11}{12}$	1

1) Tegn grafen for $F(x)$.
2) Find den tilsvarende tæthedsfunktion $f(x)$ og tegn grafen.
3) Beregn middelværdien $E(X)$, variansen $V(X)$ og spredningen $\sigma(X)$.

Opgave 3.4

Givet følgende fordelingsfunktion $F(x)$:

x	$x<0$	$0\le x<4$	$4\le x<6$	$6\le x<12$	$12\le x<18$	$18\le x$
$F(x)$	0	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{3}{4}$	1

1) Tegn grafen for $F(x)$.
2) Find den tilsvarende tæthedsfunktion $f(x)$ og tegn grafen.
3) Beregn middelværdien $E(X)$, variansen $V(X)$ og spredningen $\sigma(X)$.

Opgave 3.5

I "Spillet om sekseren" kastes 3 terninger.
Reglerne er følgende:

Antal seksere	3	2	1	0
Gevinst (i kr)	10	4	2	-2

a) Angiv sandsynlighedsfordelingen for den stokastiske variabel X, som angiver gevinsten.
b) Beregn middelværdi E(X) og spredning $\sigma(X)$.
c) Tre spillere A, B og C spiller spillet i løbet af en aften.
A spiller 80 gange , B spiller 50 gange og C spiller 20 gange
Beregn den forventede gevinst for de 3 spillere.

Opgave 3.6

Fra en beholder med 8 enheder hvoraf 2 er defekte udtages uden tilbagelægning en stikprøve på 4 enheder.
Lad X være antal defekte i stikprøven.
a) Angiv sandsynlighedsfordelingen for X.
b) Beregn middelværdi $E(X)$ og spredningen $\sigma(X)$.

4 BINOMIALFORDELING

4.1. Indledning

Næst efter normalfordelingen er binominalfordelingen nok den fordeling der har flest anvendelser.

4.2. Definition og beregning

Binomialfordelingen benyttes som model for antallet af "succeser" ved n uafhængige gentagelser af et eksperiment, som hver gang har samme sandsynlighed p for "succes".
Problemstillingen fremgår af følgende eksempel, hvor formlen samtidig "udledes".

Eksempel 4.1. En binomialfordelt variabel.
En skytte har 15% sandsynlighed for at ramme målet.
Skytten skyder 6 gange. Hvad er sandsynligheden for at skytten har netop 2 træffere.
Lad X være antallet af træffere blandt de 6 skud
Vi ønsker at finde sandsynligheden for at finde netop 2 træffere blandt disse 6, det vil sige $P(X=2)$.
Løsning:
Lad et eksperiment være at skyde et skud.
Resultatet af eksperimentet har to udfald: træffer , forbier.
Eksperimentet gentages 6 gange uafhængigt af hinanden.
Der er en bestemt sandsynlighed for at få en træffer, nemlig $p = 0.15$.
Lad t være det udfald at få en træffer, og f være det udfald at få en forbier.
Et af de ønskede forløb med 2 træffere vil eksempelvis være *t, f, t, f, f, f, f.*
Dette forløb må have sandsynligheden
$0.15 \cdot (1-0.15) \cdot 0.15 \cdot (1-0.15) \cdot (1-0.15) \cdot (1-0.15) = 0.15^2 \cdot (1-0.15)^4$.
Et andet gunstigt forløb kunne være *f, f, t, f, t, f* med sandsynligheden
$(1-0.15) \cdot (1-0.15) \cdot 0.15 \cdot (1-0.15) \cdot 0.15 \cdot (1-0.15) = 0.15^2 \cdot (1-0.15)^4$
Vi ser, at alle gunstige forløb har samme sandsynlighed.
Antal forløb må være lig antal måder man kan placere to *t*'er på 6 tomme pladser (eller antal måder man kan tage 2 kugler ud af en mængde på 6). Dette ved vi kan gøres på K(6,2) måder.
Vi får følgelig, at $p = K(6,2) \cdot 0.15^2 \cdot (1-0.15)^4 = 0.1762 = \underline{\underline{17.62\%}}$

◆

I eksemplet har vi "udledt" den såkaldte **binomialfordeling**, som er defineret på følgende måde:

DEFINITION af binomialfordeling.
1) Lad et tilfældigt eksperiment have 2 udfald "succes" og "fiasko"
2) Lad eksperimentet blive gentaget n gange uafhængigt af hinanden, og lad sandsynligheden for succes være en konstant p

Lad X være antallet af succeser blandt de n gentagelser
Der gælder da: $P(X=x) = K(n,p) \cdot p^x \cdot (1-p)^{n-x}$ for $x \in \{0,1,2,\ldots,n\}$
X siges at være binomialfordelt b (n, p).

Eksempel 4.2. Beregning af binomialfordeling
I eksempel 4.1 fandt vi, at X var binomialfordelt med $n = 6$ og $p = 0{,}15$.
$P(X = 1)$ beregnes i TI 89 på følgende måde:
CATALOG\F3\ binomPdf(6,0.15,1) Resultat: 0.3993

Binomialfordelingen anvendes sædvanligvis ved kvalitetskontrol. Man udtager her en stikprøve på n elementer ud fra en stor mængde på N elementer.

Eksempel 4.3. Kvalitetskontrol
En producent fabrikerer komponenter, som sælges i æsker med 600 komponenter i hver. Som led i en kvalitetskontrol udtages hvert kvarter tilfældigt en æske produceret indenfor de sidste 15 minutter, og 25 tilfældigt udvalgte komponenter i denne undersøges, hvorefter det foregående kvarters produktion godkendes, såfremt der højst er én defekt komponent i stikprøven.
Hvor stor er sandsynligheden (acceptsandsynligheden) p for at få partiet godkendt, hvis æsken indeholder i alt 10 defekte komponenter?

Lad os antage, der er M defekte elementer i den store mængde på N elementer
Hvis kvalitetskntrollen sker ved at man tager et emne op, undersøger det og lægger emnet tilbage, så var der jo en fast sandsynlighed på $p = \frac{M}{N}$ for at få en defekt. I et sådant tilfælde er fordelingen derfor binomialfordelt $b(n,p)$.
Sædvanligvis lægger man ikke lelmenterne tilbage, men beholder dem oppe. I så tilfælde er fordelingen hypergeometrisk.
Hvis man udtager en lille stikprøve af størrelsen n *af* en stor mængde af størrelsen N, vil sandsynligheden for at få en defekt ikke ændre sig meget hvad enten man lægger tilbage eller ej. For de fleste anvendelser kan man derfor med en passende nøjagtighed erstatte den hypergeometriske fordeling med binomialfordelingen, hvis stikprøvestørrelsen n er mindre end eller lig 10% af partistørrelsen N ($\frac{n}{N} \le \frac{1}{10}$).

Eksempel 4.3. fortsat)
Hvor stor er acceptsandsynligheden p, hvis æsken indeholder i alt 10 defekte komponenter.
Løsning:
Lad X være antallet af defekte blandt de 25 komponenter
Vi har: $p = P(X \le 1)$
Da $\frac{n}{N} = \frac{25}{600} < \frac{1}{10}$ kan approksimeres med binomialfordelingen $b\left(25, \frac{10}{600}\right)$.
TI 89: CATALOG\F3\binomCdf(25, 10/600, 0,1) =0.9353 = 93,53%

Middelværdi og spredning for binomialfordeling *b(n,p)*

Binomialfordelingen har middelværdien $\mu = n \cdot p$ *og spredningen* $\sigma = \sqrt{n \cdot p \cdot (1-p)}$.

Heraf fås (ved division med n), at p har spredningen $\sigma(p) = \sqrt{\frac{p \cdot (1-p)}{n}}$.

Et bevis vil ikke blive foretaget her.

Eksempel 4.4: Hypergeometrisk approksimeret med binomialfordeling .

Ifølge et teleselskabs opgørelse ser 70% af husstandene i en kommune med 1000 husstande fjernsyn via en parabolantanne.

En repræsentativ stikprøve på 15 husstande udtages.

Lad X = antal husstande med parabol ud af 15

Da man må antage, at man ikke spørger den samme husstand to gange er X strengt taget hypergeometrisk fordelt. Da $\frac{n}{N} = \frac{15}{1000} < 0.1$ kan man tillade sig at approksimere med binomialfordelingen b(15, 0.70). (antallet N af indbyggere i kommunen er så stort, at sandsynligheden ikke ændrer sig, fordi man har udtaget op til 15 husstande).

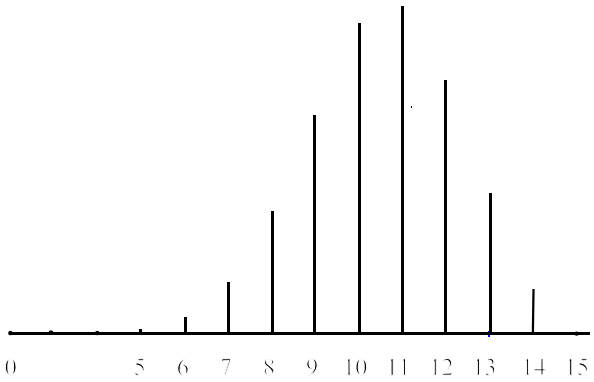
1) Tegn sandsynlighedsfunktionen for X (idet X antages binomialfordelt)
2) Beregn middelværdi og spredning for X

Løsning:

1) Da X er en "diskret" variabel, der kun antager hele værdier tegnes et stolpediagram.
   Vi beregner værdierne ved at benytte TI 89, eksempelvis
   $P(X = 9)$ = binomPdf(15,0.7,9) = 0,147
   Vi får følgende tabel:

x	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
P(X=x)	1,4E-08	5,0E-07	8,2E-06	8,3E-05	5,8E-4	3.0E-3	1,2E-2	3,5E-2	0,081	0,147	0,206	0,219	0,170	0,092	0,031	4,7E-3

   Af tabellen og af nedenstående stolpediagram ses, at vi har de største værdier sandsynligheder for x =10 og x = 11 svarende til at 70% af 15 er 10.5, og at fordelingen er nogenlunde symmetrisk omkring middelværdien 10.5.



Stolpediagram for binomialfordelingen

2) $\mu = n \cdot p = 15 \cdot 0.7 = \underline{\underline{10.5}}$ og $\sigma = \sqrt{n \cdot p \cdot (1-p)} = \sqrt{15 \cdot 0.7 \cdot (1-0.7)} = \underline{\underline{1.77}}$

4.3. Konfidensinterval for *p*

I aviser, TV m.m. optræder utallige opinionsundersøgelser og markedsundersøgelser, hvor man spørger en forhåbentlig repræsentativ stikprøve om deres mening.
Resultaterne er naturligvis usikre, men sjældent fortælles der om hvor stor usikkerheden er.
Følgende eksempel illustrerer dette.

Eksempel 4.5. Opinionsundersøgelse.
Ved valget i 2004 stemte 25.9% af vælgerne på socialdemokraterne.
I en opinionsundersøgelse svarede 1035 vælgere på spørgsmålet om hvilket parti det var mest sandsynligt de ville stemme på hvis der var valg i morgen.

a) Hvis 24.8% svarede, at de ville stemme på Socialdemokraterne, viser det så, at partiet er gået tilbage?
b) Hvis 23% svarede at de ville stemme på Socialdemokraterne, viser det så, at partiet er gået tilbage?

Løsning:

a) Idet 25.9% af 1035 er ca. 268, og 24.8% af 1035 er ca. 257.
Lad X = antal vælgere der svarer, at de vil stemme på socialdemokraterne ud af 1035 vælgere.
X antages binomialfordelt med n = 1035 og p ukendt.
Under forudsætning af at partiet har samme tilslutning som ved valget vil vi beregne sandsynligheden for at man ved opinionsundersøgelsen får 257 stemmer eller færre.
$P(X \le 257)$ =binomCdf(1035, 0.259,0,257) = 0.2275 = 22.75%
Hvis socialdemokraterne har samme tilslutning som ved valget er der altså ca. 23% sandsynlighed for at man ved en opinionsundersøgelse ville få samme resultat, dvs. at 257 (eller færre) ville sige, de ville stemme på partiet. Man kan derfor ikke med rimelig fastslå, at denne opinionsundersøgelse viser, at partiet er gået tilbage i tilslutning i forholdet til valget.
b) Hvis socialdemokraterne ved opinionsundersøgelsen kun havde fået 23% af stemmerne , svarende til ca. 238 stemmer, så finder vi tilsvarende, at
$P(X \le 200)$ =binomCdf(1035, 0.259,0,238) = 0.017 = 1.7%
Nu er der kun 1.7% sandsynlighed for at man ville få et sådant resultat, hvis tilslutningen var uændret, så umiddelbart må man konkludere, at tilslutningen er faldet.

◆

Som det fremgår af eksempel 4.5 er spørgsmålet om at tilslutningen er faldet, afhængig af hvor sikker man vil være.
Hvis man i en opinionsundersøgelsen har fået, at 24% af 1035 vælgerne har sagt. de vil stemme på socialdemokraterne, så vil man sædvanligvis ønske at angive et "usikkerhedsinterval" symmetrisk om de 24%, som angiver, at den "sande" tilslutning p til partiet med 95% sikkerhed ligger indenfor dette interval. Vil man være mere sikker, så kan man jo i stedet vælge 99% eller 99.9%, men da regningerne i princippet er de samme, vil vi i det følgende vælge 95%.

Et interval, hvor man vil være 95% sikker på at den sande værdi p ligger indenfor intervalgrænserne , kaldes et 95% konfidensinterval.

Vi har tidligere nævnt, at såfremt en fordeling er nogenlunde symmetrisk omkring middelværdien μ, så vil ca. 95% af alle værdier ligge indenfor $[\mu - 2\cdot\sigma; \mu + 2\cdot\sigma]$

For binomialfordelingen $b(n,p)$ gælder, at den har middelværdien $\mu = n \cdot p$ og spredningen $\sqrt{np(1-p)}$.

Endvidere er fordelingen rimelig symmetrisk om middelværdien når blot middelværdien ikke ligger for tæt ved 0 eller n.

Vi har nu $\mu \pm 2\sigma = np \pm 2\sqrt{np(1-p)}$

Divideres med n haves $p \pm 2 \cdot \sqrt{\frac{p(1-p)}{n}}$

Idet man kan vise, at tallet 2 mere præcist er 1.96, kan de foregående betragtninger begrunde følgende formel (som ikke bevises her)

95% konfidensinterval for binomialfordelt variabel p.

Lad der i en stikprøve på n være x Sucesser, og lad $\hat{p} = \frac{x}{n}$.

Forudsat, at $n \cdot \hat{p} \cdot (1-\hat{p}) \geq 10$ *kan et 95% konfidensinterval for p beregnes af formlen*

$$\hat{p} - 1.96 \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}} \leq p \leq \hat{p} + 1.96 \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}} \qquad (1)$$

Hvis vi i stedet ønsker et 99% konfidensinterval er den eneste ændring, at man erstatter 1.96 med 2.576.

Generelt gælder, at for et $\beta \cdot 100$ % konfidensinterval er den eneste ændring af formlen, at 1.96 erstattes af tallet invNorm((1+ β)/2,0,1).

Eksempelvis for et 95% konfidensinterval er invNorm((1+0.95)/2,0,1) = 1.96.

Eksempel 4.6 Beregning af 95% konfidensinterval

Lad os i fortsættelse af eksempel 4.5 antage, at af 1035 vælgere har 248 (ca. 24 %) sagt de vil stemme på socialdemokraterne.

Angiv på dette grundlag et 95% konfidensinterval for socialdemokraternes stemmeandel p .

Løsning:

X= antal vælgere der vil stemme på socialdemokraterne

X er binomialfordelt med n = 1035, p ukendt

$$\hat{p} = \frac{248}{1035} = 0.24$$

Da $n \cdot \hat{p} \cdot (1-\hat{p}) = 1035 \cdot 0.24 \cdot (1-0.24) = 188 > 10$ kan formlen (1) anvendes

$$\hat{p} \pm 1.96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.24 \pm 1.96\sqrt{\frac{0.24(1-0{,}24)}{1035}} = 0.24 \pm 0.026$$

95% konfidensinterval $0.24 - 0.026 \leq p \leq 0.24 + 0.026 \Leftrightarrow \underline{\underline{0.214 \leq p \leq 0{,}266}}$

TI-Nspire benytter netop denne formel i sin beregning, dvs. man skal altid her først undersøge om forudsætningen er opfyldt.

Statistik ▶ Konfidensintervaller ▶ z-interval for en andel ▶ Udfyld menu ▶ ENTER

Menuen udfyldes med x: 248 n: 1035 C-level: 0.95 ENTER

zInterval_1Prop 248,1035,0.95: **stat.results** ▸

"Titel"	"z–interval for en andel"
"CLower"	0.213609
"CUpper"	0.265618
"p̂"	0.239614
"ME"	0.026005
"n"	1035.

Resultat: C Int : [0.2136 ; 0.2656]
Vor konklusion er følgelig, at man først kan sige, at opinionsundersøgelsen (med 95% sikkerhed) viser en ændret tilslutning til socialdemokraterne, hvis valgresultatet i 2001 lå udenfor intervallet 21.4% til 26.6% . ◆

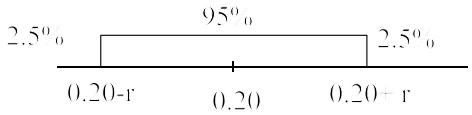
Hvis betingelsen (stikprøvestørrelsen *n* er for lille) kan man eventuelt benytte følgende generelle (men også mere besværlige) metode.

Eksempel 4.7. Beregning af konfidensinterval hvis betingelserne ikke er opfyldt
I forbindelse med et reklamefremstød ønskede man at undersøge om borgerne i en mindre by havde set en bestemt reklame. Man spurgte et antal tilfældigt udvalgte husstande, og af 50 svar havde 10 set reklamen.
Opstil et 95% konfidensinterval for sandsynligheden *p* for at man har set reklamen.
Løsning:

Vi har, at $\hat{p} = \frac{10}{50} = 20\%$

Da $n \cdot \hat{p} \cdot (1 - \hat{p}) = 50 \cdot 0.2 \cdot 0.8 = 8 < 10$ er betingelsen ikke opfyldt.

Udenfor et 95% konfidensinterval ligger 5%, og af symmetrigrunde ligger der 2,5% på hver side. (jævnfør figuren)



Jo større den sande værdi p er i forhold til 0.20 jo mindre bliver sandsynligheden for at få 10 svar eller færre. Vi leder derfor i grænsen efter et $p > 0.20$, så $P(X \le 10) = 0.025$.

solve(binomCdf(50, p,0,10)=0.025,p)$|x > 0$ Resultatet blev $p = 0.337$.

Dernæst findes nedre grænse ved at lade *p* falde, indtil $P(X \ge 10) = 0.025$

Heraf fås solve(binomCdf(50, p,10,50)=0.025,p)$|x > 0$ Resultatet blev $p = 0.100$.

Konfidensinterval: [0.100; 0.337]
Bemærk, at konfidensintervallet her ikke ligger symmetrisk omkring 0.20, da binomialfordelingen ikke i netop disse situationer netop ikke er symmetrisk omkring 0.20 ◆

Opgaver

Opgave 4.1
Under en skydeøvelse viser det sig, at en premierløjtnant rammer et mål med 40% sandsynlighed. Premierløjtnanten affyrer 8 skud.
a) Find sandsynligheden for 3 træffere.
b) Find sandsynligheden for at få mindst 3 træffere.

Opgave 4.2
På flyvestationens hovedværksted har man fået oplyst, at sandsynligheden for en defekt bolt i en boltefabrikation er 0.1. Man får en forsendelse med 400 bolte
a) Hvad er sandsynligheden for, at man i en stikprøve på 12 bolte finder mindst 1 defekt bolt.
b) Hvor mange defekte bolte vil der i middel være i forsendelsen.

Opgave 4.3
Under en øvelse affyrer en officer 80 skud mod et mål. På grund af meget vanskelige forhold er sandsynligheden for en træffer kun 0.05 i hvert forsøg.
Hvad er sandsynligheden for at officeren opnår mindst 5 træffere.

Opgave 4.4
Idet sandsynligheden for at ramme et større mål er 0.8, affyres 225 skud med en kanon. Målet anses for ødelagt, såfremt mindst 200 skud træffer det
Find sandsynligheden for at målet ødelægges.

Opgave 4.5
Når SOS har fået anvist erhvervspraktikanter, har det desværre vist sig, at kun 60% af de anviste skoleelever dukker op. Forud for årets praktik har SOS meddelt, at det kun er muligt at gennemføre praktiktjenesten, hvis der dukker mindst 12 praktikanter op.
Skoleelevernes "dukken op" er uafhængig af hinanden.
a) Hvad er sandsynligheden for at SOS kan oprette et hold, hvis praktiktjenesten anviser 15 skoleelever til SOS?
b) Hvor mange elever skal praktiktjenesten anvise, hvis der skal være 95% sandsynlighed for at kurset oprettes.

Opgave 4.6
Man udskifter i øjeblikket en ældre model fragmenteringsvest med en nyere. I lejrens depot udgør den ældre model 5% .
Ved en øvelse udleveres hurtigt og ganske tilfældigt 62 fragmenteringsveste til en deling. Hvad er sandsynligheden for, at flere end 5 fra delingen får udleveret en gammel model.

Opgave 4.7
En tipskupon har 13 kampe med 3 mulige tegn - 1, x og 2 - for hver kamp. En person bestemmer tegnet, der skal sættes for hver kamp, ved tilfældig udtrækning af en seddel fra 3 sedler med tegnene henholdsvis 1, x og 2. Angiv sandsynligheden for, at personen opnår netop 8 rigtige tippede kampe på sin kupon.

Opgave 4.8
Blandt familier med 3 børn udvælges 50 familier tilfældigt. Angiv sandsynligheden for, at der i mindst 8 af disse familier udelukkede er børn af samme køn.

Opgave 4.9
I en urne er der et meget stort antal kugler, hvoraf de 70% er sorte. Fra urnen tages en stikprøve på 10 kugler. Find sandsynligheden for, at der i stikprøven er:
1) 10 sorte kugler
2) 6, 7 eller 8 sorte kugler
3) Mindst 7 sorte kugler

Opgave 4.10
Ved et køb af 100000 plastikbægre aftaltes med leverandøren, at det skal være en forudsætning for købet, at partiet godkendes ved en stikprøvekontrol.
Kontrollen udøves ved, at 100 bægre udtages tilfældigt af partiet og kontrolleres. Partiet godkendes, såfremt ingen af de 100 bægre er defekte.
Beregn sandsynligheden for, at partiet godkendes, hvis det i alt indeholder 250 defekte bægre.

Opgave 4.11
I et elektrisk specialapparat indgår 30 komponenter, som hver er indkapslet i et heliumfyldt hylster. Beregn, idet sandsynligheden for, at et komponenthylster lækker, er 0.2%, sandsynligheden for, at mindst ét af de 30 komponenthylstre lækker.

Opgave 4.12
Det er oplyst, at der for en given vaccine er 80% sandsynlighed for, at den ved anvendelse har den ønskede virkning.
På et hospital foretoges vaccination af 100 personer med den pågældende vaccine.
Beregn sandsynligheden for, at 15 eller færre af de foretagne vaccinationer er uden virkning.

Opgave 4.13
En fabrikant får halvfabrikata hjem i partier på 200000 enheder. Fra hvert parti udtages en stikprøve på 100 enheder og antallet af fejlagtige blandt disse noteres.
Hvis dette antal er mindre end eller lig med 2, accepteres hele partiet; i modsat fald undersøges partiet yderligere.
1) Hvad er sandsynligheden for, at et parti med en fejlprocent på 1 vil blive yderligere undersøgt.
2) Hvor stor er sandsynligheden for, at et parti med en fejlprocent på 5 vil blive accepteret.

Opgave 4.14
Ved en fabrikation af plastikposer leveres disse i æsker med 100 poser i hver. Ved en godkendelseskontrol af et parti plastikposer udtages og undersøges en tilfældigt udtaget æske, og partiet godkendes, såfremt æsken højst indeholder én defekt pose.
Vi antager, at den løbende produktion af poser er således, at hver produktion med sandsynligheden 2% giver en pose, der er defekt.
Hvor stor er sandsynligheden for, at partiet under disse omstændigheder accepteres?

Opgave 4.15
En producent af billigt plastiklegetøj får mange klager over at en bestemt type legetøj er defekt ved salget. Legetøjet sælges til butikkerne i kasser på 10 stk., og som et led i en kvalitetetskontrol udtages 100 kasser og antallet x af defekt legetøj optaltes. Følgende resultater fandtes:

x	0	1	2	3	4	5	6
Antal kasser	34	38	19	6	2	0	1

Lad p være sandsynligheden for at få et defekt stykke legetøj.
1) Find et estimat $\tilde{p}$ for p.
2) Angiv et 95% konfidensinterval for p.
3) Lad X være antal defekte i en kasse på 10 stykker legetøj, og antag at X er binomialfordelt $b(10, \tilde{p})$. Beregn hvor mange af de 100 kasser, der kan forventes at have $x = 2$ defekte.

Opgave 4.16
I rapporten "Analyse af elevkampagnen 2006" udarbejdet af "Forsvarets rekruttering" returnerede 604 personer et udsendt spørgeskema.
På side 10 er en opgørelse over hvilke medier der var udslagsgivende for materialebestilling.
Der påstås side 7, at den usikkerhed der knytter sig til målingerne er $\pm 3.5\%$
Heraf fremgår at TV-spot var udslagsgivende for p = 34%
Beregn et 95% konfidensinterval for p, og kommenter ovennævnte påstand.

Opgave 4.17
I en analyse af arbejdsgivernes tilfredshed med jobnet, svarede 488 arbejdsgivere på spørgsmålet.
Det viste sig, at kun 5% var utilfredse med jobnet.
Beregn et 95% konfidensinterval for $p = 0.05$.

Opgave 4.18
I en analyse blev 428 arbejdsgivere spurgt om hvilke jobtyper de annoncerede på jobnet.
Det viste sig, at kun 7% benyttede jobnet til at annoncere efter ledere.
Beregn et 95% konfidensinterval for $p = 0.07$

Opgave 4.19
En ny behandling af cancer forventes at give bedre overlevelseschancer end den hidtidige behandling. 120 patienter prøvede den nye behandling, og af disse overlevede 82 i mere end 5 år.
Idet antallet af overlevende patienter antages at være binomialfordelt, skal man
1) Angive et estimat for sandsynligheden p for at overleve i 5 år ved den nye behandling.
2) Angive et 95% konfidensinterval for p.

5 Deskriptiv Statistik

5.1 Indledning

Statistik kan lidt løst sagt siges, at være en samling metoder til at opnå og analysere data for at træffe afgørelser på grundlag af dem.
Statistik er et uundværligt værktøj til at træffe beslutninger, men kan naturligvis som alt andet også misbruges, bevidst eller ubevidst. Beslutninger der kan basere sig på tal (statistik), får stor troværdighed. Det kan bevirke at man slår sin "sunde fornuft" fra. Selv den bedste statistiske teori er værdiløs, hvis tallene man bygger på ikke er troværdige, eller relevante, og det er derfor ikke så mærkeligt, at en kendt politiker engang udtalte:"Der findes 3 slags løgn: løgn, forbandet løgn og statistik".

Ved **populationen** forstås hele den gruppe man er interesseret i. Eksempelvis hvis det drejer sig om folketingsvalg i Danmark, så er populationen alle stemmeberettigede personer i Danmark . Ved en **stikprøve** forstås en delmængde af populationen. Før et folketingsvalg udtager et opinionsinstitut således en stikprøve på eksempelvis 1000 vælgere.

Der er to grundlæggende anvendelser af statistik:
1) Deskriptiv statistik, hvor man sammenligner og beskriver data.
   Eksempelvis kunne man sammenligne hvormange personer, der stemte på partierne ved sidste og næstsidste valg.
2) "inferens" statistik , hvor man ved anvendelse af statistiske metoder søger at slutte (informere) fra en stikprøve til hele proportionen.
   Eksempelvis før et folketingsvalg på basis af en stikprøve på 1000 personer der bliver spurgt om hvem de vil stemme på give en prognose for den forventede mandatfordeling for hele landet (populationen)
   Her vil det være nødvendigt med at kende nogle statistiske metoder til eksempelvis at vide hvor stor en (repræsentativ) stikprøve man skal udtage for at usikkerheden på resultatet er under 5%

5.2.Grafisk beskrivelse af data

I den **deskriptive statistik** (eller beskrivende statistik) beskrives de indsamlede data i form af tabeller, søjlediagrammer, lagkagediagrammer, kurver samt ved udregning af centrale tal som gennemsnit, typetal, spredning osv.
Kurver og diagrammer forstås lettere og mere umiddelbart end kolonner af tal i en tabel. Øjet er uovertruffet til mønstergenkendelse ("en tegning siger mere end 1000 ord").
Vi vil i dette afsnit benytte programmet Excel, da det bedre end en lommeregner kan beskrive data grafisk.

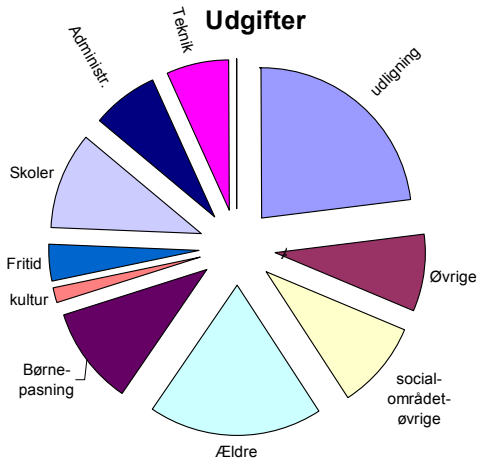
Kvalitative data

Hvis der er en naturlig opdeling af talmaterialet i klasser eller kategorier siges, at man har kategorisk eller kvalitative data .
Alle spørgeskemaundersøgelser, hvor man eksempelvis bliver bedt om at sætte kryds i nogle rubrikker "meget god" , god, acceptabel osv. er af denne type.

Til illustration af disse data bruges sædvanligvis lagkagediagrammer eller søjlediagrammer

Eksempel 5.1 Lagkagediagram

Et eksempel ses overfor, hvor et lagkagediagram søger at give et anskueligt indtryk af hvordan en kommunes udgifter fordeler sig på de forskellige områder.



I Excel opskrives

Udligning	23,1
øvrige	8,4
Socialområdet,øvrige	9,4
Ældre	18,6
Børnepasning	10,4
Bibliotek	1,9
fritid	3,8
Skoler	10,5
Administration	7,3
Teknik,anlæg	6,6

Excel-ordrer:

2003: Marker udskriftsområde ⇒ Vælg på værktøjslinien "Guiden diagram" ⇒ Cirkel ⇒ Marker ønsket figur ⇒ Næste ⇒ Navn på kategori ⇒ Udfør

2007: Marker udskriftsområde ▶ Vælg på værktøjslinien "Indsæt" ▶ Cirkel ▶ Marker ønsket figur ◆

Eksempel 5.2 (kvalitative data)

Følgende tabel angiver mandattallet ved de to sidste folketingsvalg.

Partier		A	B	C	F	K	O	V	Ø
Mandater	2001	52	9	16	12	4	22	56	4
	2005	47	17	18	11	0	24	52	6

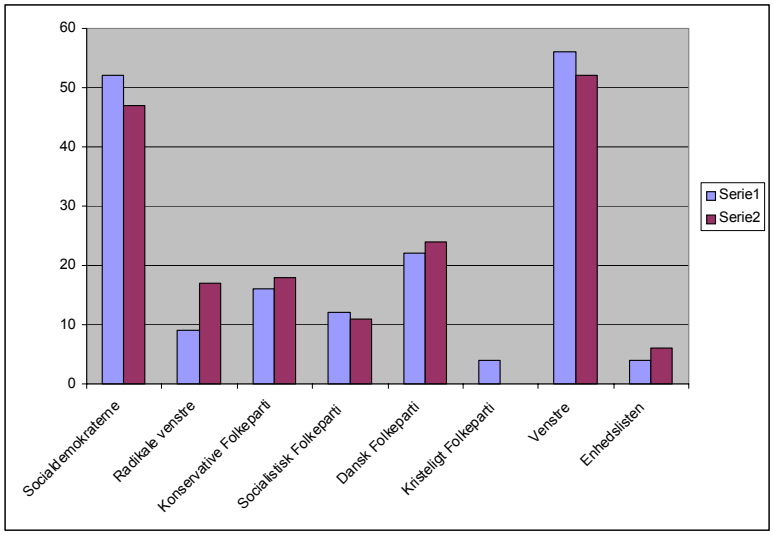
A = Socialdemokraterne, B = Radikale venstre, C = Konservative folkeparti , F = Socialistisk folkeparti, K = Kristendemokraterne, O = Dansk Folkeparti, V = Venstre, Ø = Enhedslisten

Et søjlediagram fås i Excel ved at opskrive

A	B	C	F	K	O	V	Ø
52	9	16	12	4	22	56	4
47	17	18	11	0	24	52	6

Excel 2003: Vælg på værktøjslinien "Guiden diagram" ▶ Søjle ▶ Marker ønsket figur ▶ Næste ▶ marker udskriftsområde ▶ Næste ▶ Næste ▶ Udfør

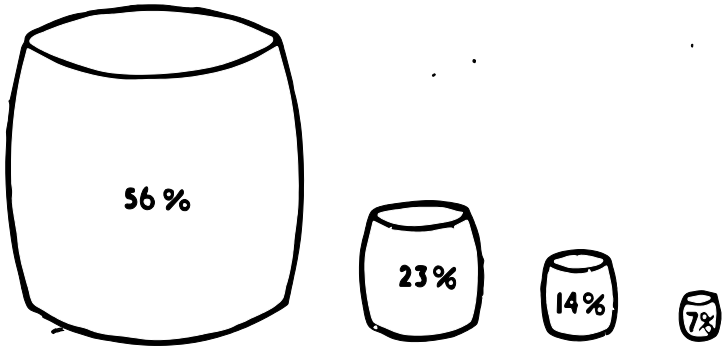
Excel 2007: Marker udskriftområde ▶ Vælg på værktøjslinien "Indsæt" ▶ Søjle ▶ Marker ønsket figur



Fordelen ved en grafisk fremstilling er, at de væsentligste egenskaber ved data opnås hurtigt og sikkert. Men netop det, at figurer appellerer umiddelbart til os, gør at vi kan komme til at lægge mere i dem, end det som tallene egentlig kan bære. Eksempelvis viser forsøg, at i lagkagediagrammer, hvor man skal sammenligne vinkler (eller arealer), da vil denne sammenligning afhænge noget af i hvilken retning vinklens ben peger.
Nedenstående eksempel viser hvordan en figur kan være misvisende uden direkte at være forkert.

Eksempel 5.3. Misvisende figur

Tønderne i figuren nedenfor skal illustrere hvordan osteeksporten fordeler sig på de forskellige verdensdele. Den giver imidlertid et helt forkert indtryk. Det er højderne på tønderne der angiver de korrekte forhold, men af tegningen vil man tro, at det er rumfangene af tønderne. De 3 små tønder kan umiddelbart være flere gange indeni den store tønde, men det svarer jo ikke til talforholdene.



Kvantitative data

Kvantitative data er data, hvor registreringen i sig selv er tal, der angiver en bestemt rækkefølge.

Eksempel 5.4. Histogram , sideafvigelse ved skydning.
Man har 100 gange målt sideafvigelsen ved skydning med maskingevær.
Resultaterne (som kan findes på adressen www.larsen-net.dk) var følgende:

33.22	21.75	5.60	4.70	9.19	11.03	-0.8	-19.01	11.08	10.91	6.93	14.6
-11.5	2.19	14.47	11.27	22.06	11.81	19.53	13.25	6.1	1.14	14.1	-4.23
9.33	14.26	-4.16	20.88	-13.29	-6.53	-3.03	0.49	13.08	3.7	-0.56	-0.36
22.29	9.01	21.49	5.1	17.88	2.68	5.23	2.81	-5.64	11.63	3.21	-0.19
18.67	17.01	-6.34	21.6	11.26	9.63	-5.97	6.42	14.65	-0.77	0.31	-0.43
2.26	6.14	12.56	11.81	11.76	23.92	4.66	23.98	4.81	26.44	4.67	21.38
-0.52	5.51	-24.44	-5.0	13.95	-6.66	10.63	10.00	-1.69	-0.37	7.59	24.22
24.16	30.22	-11.84	14.45	-12.27	18.94	0.85	9.93	8.89	9.64	-3.28	16.27
16.63	5.87	4.35	6.7								

Giv en grafisk beskrivelse af disse data.

Løsning:
I dette tilfælde, hvor vi er interesseret i at få et overblik over tallenes indbyrdes størrelse er det fordelagtigt at tegne et **histogram**.
Et histogram ligner et søjlediagram, men her gælder, at antallet af enheder i hver søjle repræsenteres ved søjlens areal (histo er græsk for areal). Man bør så vidt muligt sørge for at grupperne er lige brede, da antallet af enheder så svarer til højden af søjlen.
Excel og TI 89 kan ummiddelbart tegne et histogram, men af hensyn til det følgende forklares, hvordan man bestemmer intervalopdeling m.m.
Først findes det største tal x_{max} og det mindste tal x_{min} i materialet og derefter beregne **variationsbredden** x_{max} - x_{min}. Vi ser, at største tal er 33.22 og mindste tal er -24.44 og variationsbredden derfor 33.22 -(-24.44) = 57.66.
Dernæst deles tallene op i et passende antal intervaller (klasser). Som det første bud vælges ofte et antal nær $\sqrt{n}$. Da $\sqrt{100} = 10$ vælges ca. 10 klasser. Da $\frac{57.66}{10} \approx 5.8$ deler vi op i de klasser, der ses af tabellen. Dette giver 11 intervaller. Vi tæller op hvor mange tal der ligger i hvert interval (gøres nemmest ved at starte forfra og sæt en streg i det interval som tallet tilhører).

Klasser		Antal *n*
]-24.5 ; -18.7]	//	2
]-18.7 ; -12.9]	/	1
]-12.9 ; - 7.1]	///	3
]-7.1 ; - 1.3]	//////////	11
]-1.3 ; 4.5]	//////////////////	19
]4.5 ; 10.3]	//////////////////////	23
]10.3 ; 15.1]	//////////////////	20
]15.1 ; 21.9]	//////////	12
]21.9 ; 27.7]	///////	7
]27.7 ; 33.5]	//	2

Excel

Data indtastes i eksempelvis søjle A1 til A100 (data findes på adressen www.larsen-net.dk)

2003: Vælg "Funktioner", Dataanalyse, Histogram

2007: Vælg "Data", Dataanalyse, Histogram

I den fremkomne tabel udfyldes "inputområdet" med A1:A100 og man vælger "diagramoutput"..

1) Trykkes på OK fås en tabel med hyppigheder, og en figur, hvor intervalgrænserne er fastlagt af Excel.
2) Ønsker man selv at bestemme grænserne, skal man også udfylde intervalområdet. Dette gøres ved at skrive de øvre grænser i en søjle (f.eks. i B1 -18.7, i B2 -12.9 osv.) og så skrive B1:B11 i inputområdet

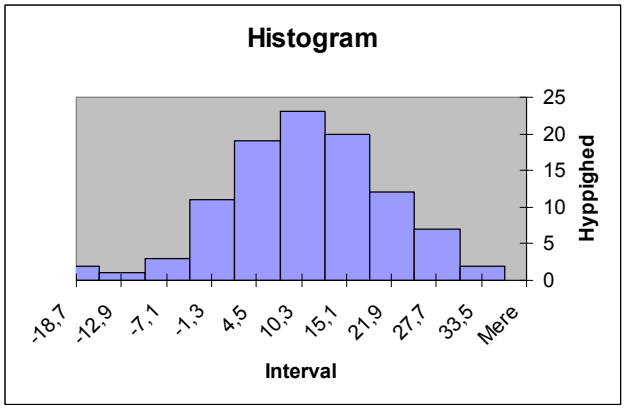
Nedenstående figurer er blevet gjort lidt "pænere" ved

a) cursor på en søjle ▶ tryk højre musetast ▶ formater dataserie ▶ indstilling ▶ mellemrumsbredde = 0 ▶ ok

TI 89

Vælg APPS\ Data/Matrix\New\Indtast data i eksempelvis søjle c1\F2: Plot Setup\F1: Define\Sæt "Plot Type" til Histogram\ENTER\HOME\Graph

Der fremkommer så i Excel følgende tegning:



Det ses, at de fleste målinger ligger fra ca. -1.3 til ca. 15.1 og så falder hyppigheden nogenlunde symmetrisk til begge sider.
Man regner normalt med, at resultaterne af forsøg, hvor man har foretaget målinger (hvis man lavede nok af dem) har et sådant klokkeformet histogram og at fordelingen er normalfordelt (beskrives nærmere i næste kapitel)

Sumpolygon

Ud over at tegne histogrammer for en stikprøve er det også ofte nyttigt, at betragte en sumpolygon for en stikprøve.

Eksempel 5.5 Sumpolygon

Lad os igen betragte de 100 sideafvigelser i eksempel 1.5.
Vi foretager nu en opsummering(kaldes kumulering), og derefter beregnes ved division med 100 (antal sideafvigelser) tallene i % af det totale antal
Derved fremkommer følgende tabel:

Klasser	Antal	Sum	Kumuleret relativ hyppighed
]-24.5 ; -18.7]	2	2	0.02
]-18.7 ; -12.9]	1	3	0.03
]-12.9 ; -7.1]	3	6	0.06
]-7.1 ; - 1.3]	11	17	0.17
]-1.3 ; 4.5]	19	36	0.36
]4.5 ; 10.3]	23	59	0.59
]10.3 ; 15.1]	20	79	0.79
]15.1 ; 21.9]	12	91	0.91
]21.9 ; 27.7]	7	98	0.98
]27.7 ; 33.5]	2	100	1.00

Afsættes punkterne (-18.7 , 0.02), (-12.9, 0.03) . . . (33.5, 1.00) (bemærk at x-værdierne er værdierne i højre intervalendepunkt), og forbindes de enkelte punkter med rette linier, fås den i figur 1.1 angivne sumpolygon, hvoraf man kan aflæse, at 25% af sideafvigelserne ligger under ca. 1. (kaldes 25% fraktilen eller første kvartil).

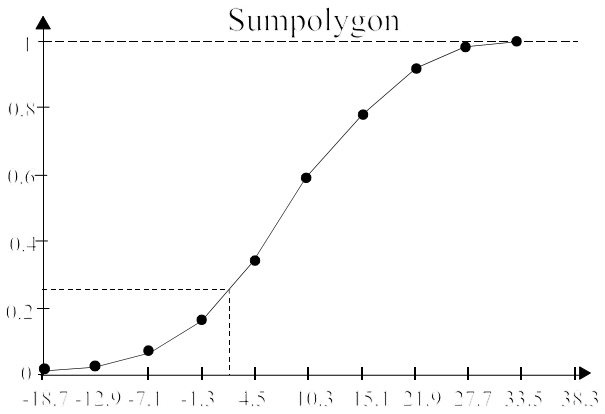


Fig 1.1 *Sumpolygon*

5.3 Karakteristiske værdier

Har man et stort talmateriale, er det nødvendigt ud over at betragte histogrammer og sunpolygoner, at give en karakteristik af dataene ved at beregne tal som kan give et mål for midterværdier og spredning.

Midterværdier

Gennemsnittet $\bar{x}$ (kaldt x streg) beregnes på sædvanlig måde. Eksempelvis har tallene 2,4,5,9 gennemsnittet $\bar{x} = \frac{2+4+5+9}{4} = 5$

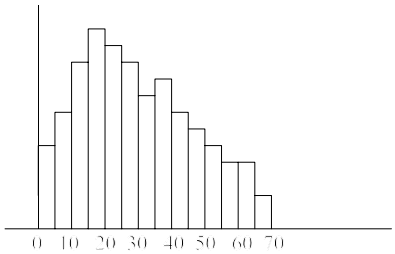
TI 89: CATALOG \mean({2,4,5,9})

Median: Medianen beregnes på følgende måde:
1) Observationerne ordnes i rækkefølge efter størrelse.
2a) Ved et ulige antal observationer er medianen det midterste tal
2b) Ved et lige antal er medianen gennemsnittet af de to midterste tal.
Eksempel: Observationer 6, 17, 7, 13, 5, 2. Ordnet i rækkefølge: 2, 5, 6, 7 13, 17. Median 6,5
TI 89: CATALOG \median({6,17,7,13,5,2})

Medianen kaldes også for **50% fraktilen**, fordi den brøkdel (fraktil) der ligger under medianen er ca. 50% .
1. og 3. kvartil svarer tilsvarende til at henholdsvis 25% og 75% fraktilen.

For store talmængder som eksempelvis de 100 værdier i eksempel 5.4 er det hvis man benytter TI 89 mest praktisk at vælge
APPS\Data/Matrix\indtaster de 100 tal i eksempelvis "c1"
F5: Calculation Type: Vælg "One Var", x = c1 ENTER
Udskriften består af en række statistiske størrelser hvoriblandt
gennemsnit = $\overline{x}$, spredning= sx, median=medStat, 1 kvartil = q1 og 3 kvartil=q3

Er median og gennemsnit nogenlunde lige store er fordelingen nogenlunde symmetrisk omkring middelværdien.
Er medianen mindre end gennemsnittet er der tale om en "højreskæv" fordeling som har den "lange" hale til højre.(se figuren)



At man eksempelvis i lønstatistikker[1] angives medianen og ikke gennemsnittet fremgår af følgende lille eksempel.
Lad os antage at en virksomhed har 10 ansatte, med månedslønninger ordnet efter størrelse på
20000, 21000, 22000, 23000, 24000, 25000, 26000, 27000, 28000, 100000
Gennemsnittet er her 31600, mens medianen er 24500.
Medianen ændrer sig ikke selv om den højeste løn vokser fra 100000 til 1 million, mens gennemsnittet naturligvis vokser. Medianen giver derfor en mere rimelig beskrivelse af middellønnen i firmaet.

I nævnte lønstatistik[1] er også angivet "nedre og øvre Kvartil som er henholdsvis 25% fraktilen og 75% fraktilen. Ved at angive dem får man et indtryk af, hvor stor lønspredningen er som det vil fremgå i afsnittet om spredning

Spredningsmål.

Støj

Egentlige målefejl, såsom at nogle af observationerne ikke bliver korrekt registreret, uklarheder i spørgeskemaet osv. skal naturligvis fjernes.
Derudover er der den "naturlige" variation som også kunne kaldes "ren støj" (pure error), som skyldes, at man ikke kan forvente, at to personer der på alle områder er stillet fuldstændigt ens også vil svare ens på et spørgsmål. Tilsvarende hvis man måler udbyttet ved en kemisk proces, så vil udfaldet af to forsøg ikke være ens, da der altid er en række ukontrollable støjkilder (urenheder i råmaterialer, lidt forskel på personer og apparatur osv.)
Denne naturlige variation skal naturligvis inddrages i den statistiske behandling af problemet, og dertil spiller et mål for, hvor meget tallene spreder sig naturligvis en væsentlig rolle..

[1] jævnfør statistisk årbog 2005 tabel 144

Kvartilafstand: Hvis fordelingen ikke er rimelig symmetrisk, er medianen det bedste skøn for en midterværdi, og kvartilafstanden kan være et mål for spredningen.

Eksempel 5.6. Kvartilafstand
I den tidligere omtalte lønstatistik[2] findes bl.a. følgende tal, idet de to sidste kolonner er vor bearbejding af tallene.

		Løn pr. præsteret time					
nr		gennem-snit $\bar{x}$	nedre kvartil k1	median m	øvre kvartil k3	$\frac{\bar{x}}{m}$	$\frac{k3-k1}{m}$
1	Ledelse på højt niveau	353.41	231.63	313.38	433.78	1.13	0.64
2	Kontorarbejde	196.82	158.86	186.99	222.78	1.05	0.34

Af kolonnen $\frac{\bar{x}}{m}$ ses, at for begge rækker er gennemsnittet større end medianen dvs. begge fordelinger er højreskæv, men det gælder mest for række nr. 1. Her gælder åbenbart, at nogle få forholdsvis høje lønninger trækker gennemsnittet op.
Skal man sammenligne lønspredningen i de to tilfælde, må man tage hensyn til, at medianen er meget forskellig. Man vil derfor som der er sket i sidste kolonne beregne den **relative kvartil-afstand.**
Den viser også, at lønspredningen er væsentlig mindre for række 2 end for række 1 .

Spredning (også kaldet standardafvigelse efter engelsk: standard deviation)

Spredningen på en stikprøve benævnes s.

s beregnes af formlen $s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$ hvor observationerne i en stikprøve er $x_1, x_2, \ldots, x_n$

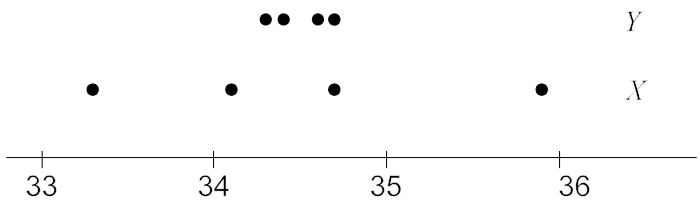
Da spredningen baserer sig på alle observationer i stikprøven (eller populationen) er den derfor (hvis fordelingen er nogenlunde symmetrisk (normalfordelt) det mest anvendte mål.

Stikprøvevariansen (eller blot variansen) er s^2 .

Eksempel 5.7: Beregning af spredning
Tallene 2,4,5,9 med $\bar{x} = 5$,har variansen
$s^2 = \frac{(2-5)^2 + (4-5)^2 + (5-5)^2 + (9-5)^2}{4-1} = \frac{26}{3} = 8.6667$ og spredningen $s = \sqrt{8.6667} = 2.9439$
TI 89: CATALOG \Variance({2,4,5,9}), CATALOG \stdDev({2,4,5,9})
Har man mange tal kan det igen betale sig at indtaste tallene i en liste.

[2] jævnfør statistisk årbog 2005 tabel 144

Anskuelig forklaring på formlen for *s*.
At formlen for *s* skulle være særlig velegnet til at angive, hvor meget resultaterne "spreder sig" (hvor megen støj der er) er ikke umiddelbart indlysende. I det følgende gives en anskuelig forklaring.
Lad os betragte 2 forsøgsvariable *X* og *Y*, hvorpå der for hver er udført en stikprøve på 4 forsøg.
Resultaterne var: *X*: 35.9, 33.3, 34.7, 34.1 med gennemsnittet $\bar{x}$ = 34.5 , og
Y: 34.3, 34.6, 34.7, 34.4 med gennemsnittet $\bar{y}$ = 34.5.



De to forsøgsvariable har samme gennemsnit, men det er klart, at *Y*-resultaterne grupperer sig meget tættere om gennemsnittet end *X*-resultaterne, dvs. *Y*-stikprøven har mindre spredning (der er mindre støj på *Y* - forsøget) end *X*-stikprøven.
For at få et mål for stikprøvens spredning beregnes resultaternes afvigelser fra gennemsnittet.

$x_i - \bar{x}$	$y_i - \bar{y}$
35.9 - 34.5 = 1.4	34.3 - 34.5 = - 0.2
33.3 -34.5 = - 1.2	34.6 - 34.5 = 0.1
34.7 - 34.5 = 0.2	34.7 - 34.5 = 0.2
34.1 - 34.5 = - 0.4	34.4 - 34.5 = -0.1

Summen af disse afvigelser er naturligvis altid 0 og kan derfor ikke bruges som et mål for stikprøvens spredning.
I stedet betragtes summen af kvadraterne på afvigelserne (forkortet SS: **S**um of **S**quares eller SAK: **S**um af **a**fvigelsernes **K**vadrat).

$$SAK_x = \sum_{i=1}^{n}(x_i - \bar{x})^2 = 1.4^2 + (-1.2)^2 + 0.2^2 + (-0.4)^2 = 3.60$$

$$SAK_y = \sum_{i=1}^{n}(y_i - \bar{y})^2 = (-0.2)^2 + 0.1^2 + 0.2^2 + (-0.1)^2 = 0.10$$

Da et mål for variansen ikke må være afhængig af antallet af forsøg, divideres med *n* - 1.
Umiddelbart ville det være mere rimeligt at dividere med *n*. Imidlertid kan det vises, at i middel bliver et skøn for variansen for lille, hvis man dividerer med *n*, mens den "rammer" præcist, hvis man dividerer med *n* - 1. Det kan forklares ved, at tallene x_i har en tendens til at ligge tættere ved deres gennemsnit $\bar{x}$ end ved middelværdien μ.

$s_x^2 = \frac{3.60}{4-1} = 1.2$ $s_y^2 = \frac{0.1}{4-1} = 0.0333$. $s_x = \sqrt{1.2} = 1.095$ og $s_y = \sqrt{0.0333} = 0.183$

Som vi forudså, er stikprøvens spredning betydelig større for *X*-resultaterne end for *Y*-resultaterne.

Frihedsgrader. Man siger, at stikprøvens varians er baseret på $f = n - 1$ **frihedsgrader**. Navnet skyldes, at kun *n* -1 af de *n* led $x_i - \bar{x}$ kan vælges frit, idet summen af de *n* led er nul. Eksempelvis ser vi af ovenstående eksempel, at der er 3 frihedsgrader, da kendskab til de første 3 led på 1.4, -1.2 og 0.2 er nok til at bestemme det fjerde led, da summen er nul.

Vurdering af størrelsen af stikprøvens spredning.
Man kan vise, at for tæthedsfunktioner med kun et maksimumspunkt gælder, at mellem $\bar{x} - 2 \cdot s$ og $\bar{x} + 2 \cdot s$ ligger ca. 89% af resultaterne, og mellem
$\bar{x} - 3 \cdot s$ og $\bar{x} + 3 \cdot s$ ligger ca. 95% af resultaterne.
For normalfordelingen er de tilsvarende tal 95% og 99%.(se figur 5.2)

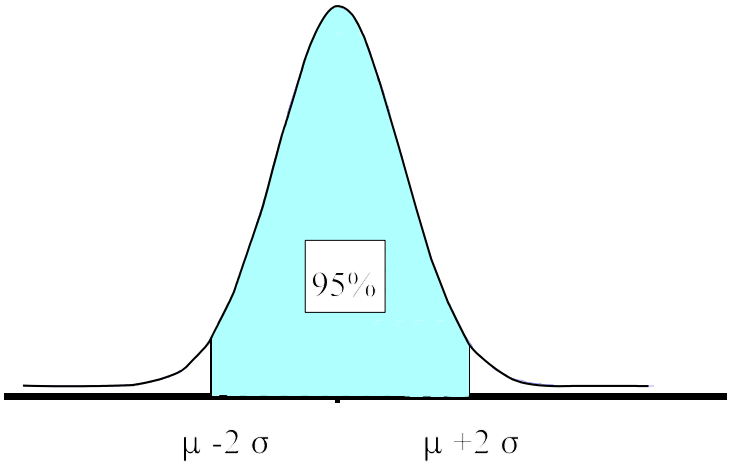


Fig. 5.2 . *Normalfordeling*

Opgaver

Opgave 5.1

Færdselspolitiet overvejede, om der burde indføres en fartgrænse på 70 km/h på en bestemt landevejsstrækning, hvor der hidtil havde været en fartgrænse på 80 km/h.

Som et led i analysen af hensigtmæssigheden af den overvejede ændring observeredes inden for et bestemt tidsrum ved hjælp af radarkontrol de forbipasserende bilers fart. Resultatet af målingerne var:

50 observationer									
64	72	82	52	60	95	86	70	63	48
50	63	35	60	77	41	47	88	62	66
59	49	55	99	65	76	76	68	51	80
75	74	64	74	62	70	85	73	93	65
98	55	85	80	78	53	96	71	84	103

1) Foretag en vurdering af, om fordelingen er nogenlunde symmetrisk (normalfordelt) ved
   a) at tegne et histogram
   b) at beregne middelværdi og median for at sw om de er nogenlunde ens.

Opgave 5.2

Til fabrikation af herreskjorter benyttes et råmateriale, som indeholder en vis procentdel uld. For nærmere at undersøge uldprocenten, måles denne i 64 tilfældigt udvalgte batch. Resultatet var (i %):

34.2	33.1	34.5	35.6	36.3	35.1	34.7	33.6	33.6	34.7	35.0	35.4	36.2	36.8	35.1	35.3
33.8	34.2	33.4	34.7	34.6	35.2	35.0	34.9	34.7	33.6	32.5	34.1	35.1	36.8	37.9	36.4
37.8	36.6	35.4	34.6	33.8	37.1	34.0	34.1	32.6	33.1	34.6	35.9	34.7	33.6	32.9	33.5
35.8	37.6	37.3	34.6	35.5	32.8	32.1	34.5	34.6	33.6	24.1	34.7	35.7	36.8	34.3	32.7

1) Foretag en vurdering af, om fordelingen er nogenlunde symmetrisk (normalfordelt) ved
   a) at tegne et histogram
   b) at beregne middelværdi og median

   Der er i datamaterialet en såkaldte outliers (en mulig fejlmåling). En sådan kan ødelægge enhver analyse. Det er i dette tilfælde tilladeligt at fjerne den, da vi går ud fra det er en fejlmåling.

2) Beregn stikprøvens relative kvartilafstand

6. Normalfordeling

6.1 Indledning

Vi har i kapitel 5 set et eksempel på en stikprøve, hvor dets histogram var (næsten) symmetrisk og "klokkeformede". Vi nævnte da, at så var den tilhørende stokastiske variabel nok "normalfordelt".
Dette er ikke tilfældigt, idet normalfordelingen er den fordeling som oftest forekommer i forbindelse med løsning af "praktiske" problemer.
Dette skyldes, at når måleresultater påvirkes af en lang række små uafhængige påvirkninger, vil observationerne være fordelt symmetrisk om en midterværdi med flest resultater tættest ved midterværdien. Måler man f.eks. vægten af syltetøj, der fyldes på en dåse af en automatisk påfyldningsmaskine, så vil denne variere på grund af mange små uafhængige og ukontrollable påvirkninger. De fleste dåsers vægt vil ligge tæt på gennemsnitsvægten, nogle vil være lidt lettere, andre lidt tungere men de vil fordele sig symmetrisk omkring middelværdien. Andelen af meget tunge dåser og meget lette dåser vil være meget lille. En sådan symmetrisk fordeling med en aftagende forekomst af observationer når vi fjerner os fra middelværdien, er netop typisk for en normalfordelt variabel.
Andre eksempler på normalfordelte variable er måling af :
rekrutters højde eller vægt, pH i ledvæsken i knæ, udbyttet af et stof A ved en kemisk proces, diameteren af en serie aksler produceret på samlebånd, udbyttet pr. hektar på hvedemarker.

6.2. Tæthedsfunktion.

Relativ hyppighed
Ved den relative hyppighed forstås hyppigheden divideret med det totale antal.

I eksempel 5.4 er den relative hyppighed for sideafvigelsen i intervallet]4.0 ; 9.3] $\frac{23}{100} = 23\%$

Man kunne sige, at "sandsynligheden" er 23% for at sideafvigelsen ligger i dette interval.
Skal man sammenligne to talmaterialer, eksempelvis sammenligne de 100-værdier i eksempel 1.5 med 200 resultater fra en anden skydebane, har det ingen mening at sammenligne hyppighederne, men derimod de relative hyppigheder, dvs. dividere hyppighederne med henholdsvis 100 og 200.

Eksempel 6.1 Tæthedsfunktion

I den følgende tabel er dels beregnet de relative hyppigheder for tallene i eksempel 5.4 dels er der af hensyn til det følgende foretaget en skalering ved at dividere den relative hyppighed med intervallængden 5.8.

Klasser	Antal n	Relativ hyppighed $\frac{n}{100}$	Skalering $\frac{n}{100 \cdot 5.8}$
]-24.5 ; -18.7]	2	0.02	0.00345
]-18.7 ; -12.9]	1	0.01	0.00172
]-12.9 ; -7.1]	3	0.03	0.00517
]-7.1 ; - 1.3]	11	0.11	0.0190
]-1.3 ; 4.5]	19	0.19	0.0328
]4.5 ; 10.3]	23	0.23	0.0400
]10.3 ; 15.1]	20	0.20	0.0345
]15.1 ; 21.9]	12	0.12	0.021
]21.9 ; 27.7]	7	0.07	0.012
]27.7 ; 33.5]	2	0.02	0.00345

Hvis man tænker sig histogrammet tegnet med de skalerede værdier i stedet for hyppighederne, så vil arealet af hver søjle være den relative hyppighed og det samlede areal være 1.

Hvis man tænker sig antallet af forsøg stiger (for eksempel ikke skyder 100 skud men måske 1 million skud), samtidig med at man øger antallet af klasser tilsvarende (til for eksempel $\sqrt{10^6} \approx 1000$) , vil histogrammet blive mere og mere fintakket, og til sidst nærme sig til en kontinuert klokkeformet kurve. For denne idealiserede kontinuerte kurve, vil arealet under kurven i et bestemt interval fra a til b være sandsynligheden for at få en værdi mellem a og b. Det samlede areal under kurven er naturligvis 1.
Man siger, at den (kontinuerte) **stokastiske** variabel X (X er her sideafvigelsen) har en **tæthedsfunktion $f(x)$** hvis graf er den ovenfor nævnte kontinuerte kurve. ◆

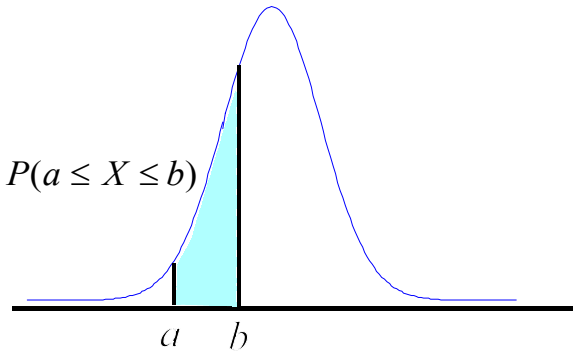
Eksempel 6.1 begrunder, at en tæthedsfunktion for en kontinuert stokastisk variabel X skal have den egenskab, at sandsynligheden for at X ligger mellem 2 værdier a og b lig med arealet under kurven[1].

Sandsynligheden for at X ligger mellem a og b skrives kort $P(a \le X \le b)$

(P står for probability)

[1] En tæthedsfunktion for en kontinuert statistisk variabel skal tilfredsstille følgende betingelser:

1) $f(x) \ge 0$, 2) $\int_{-\infty}^{\infty} f(x)dx = 1$, 3) $P(a \le x \le b) = \int_a^b f(x)dx$ for ethvert interval [a ; b]



På basis af en stikprøve på *n* tal , kunne vi regne gennemsnit $\bar{x}$ og spredning *s* ud.

Middelværdi: Kendes den stokastiske variabel *X* 's tæthedsfunktion *f*(*x*) kan beregnes en "korrekt midterværdi". Denne kaldes middelværdien for *X* og benævnes μ eller *E*(*X*) (*E* for expected).[2]

Spredning (også kaldet standardafvigelse efter engelsk: standard deviation)
Tilsvarende kan beregnes en eksakt værdi for spredningen. Denne benævnes σ eller $\sigma(X)$

Man siger kort, at gennemsnittet $\bar{x}$ er et **estimat** for μ, og "stikprøvens spredning" *s* er et **estimat** for σ.

Ofte regner man i variansen, som benævnes σ^2 eller *V*(*X*).

---

[2] **Definition: Middelværdi** $E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx$ **Varians** $\mathrm{V(X)} = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)dx$

Spredning $\sigma(X) = \sqrt{\mathrm{V}(X)}$

6.3 Tætheds- og fordelingsfunktion for normalfordeling

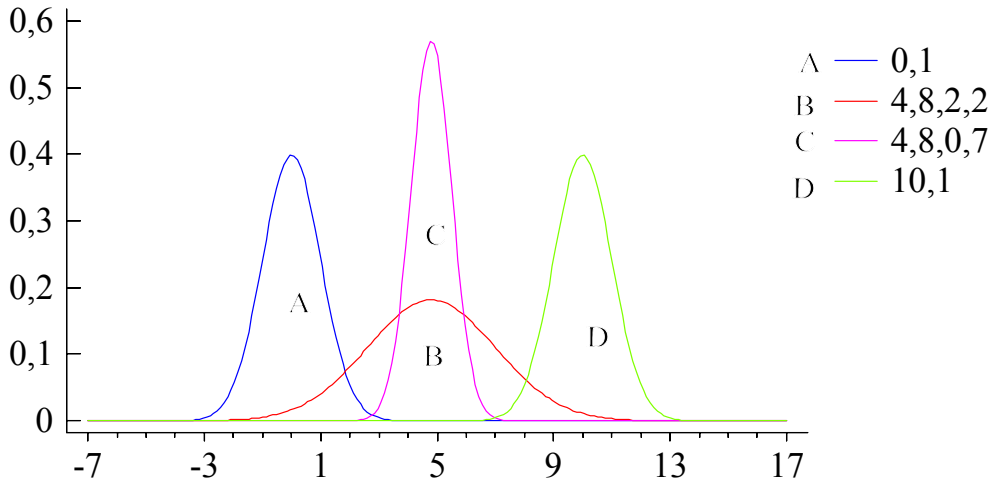
Som nævnt i indledningen er den såkaldte normalfordeling af særlig interesse.

Normalfordelingen med middelværdi μ og spredningen σ benævnes kort $n(\mu,\sigma)$.

Tæthedsfunktion.

Tæthedsfunktionen $f(x)$ [3] er særdeles kompliceret, men da den findes på mange regnemaskiner såsom TI 89, er dette ikke noget regneteknisk problem.

For at få et overblik over betydningen af μ og σ er der nedenfor afbildet tæthedsfunktionerne for normalfordelingerne $n(0\ ,\ 1)$, $n(4.8\ ,\ 2.2)$, $n(4.8\ ,\ 0.7)$ og $n(10\ ,\ 1)$.



Arealerne under kurverne er alle 1, og man ser, at "klokkeformen" bliver bred når spredningen er stor. Et interval på $[\mu-3\cdot\sigma;\mu+3\cdot\sigma]$ indeholder stort set hele sandsynlighedsmassen, og et interval på $[\mu-2\cdot\sigma;\mu+2\cdot\sigma]$ indeholder ca. 95% af sandsynlighedsmassen.

[3] Normalfordelingen har funktionsforskriften er $f(x)=\frac{1}{\sqrt{2\cdot\pi}\cdot\sigma}\cdot e^{-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2}$, $-\infty<x<\infty$

Fordelingsfunktion

Fordelingsfunktionen $F(x)$ for en kontinuert variabel X er defineret ved

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(x)dx$$

Grafen for $F(x)$ kan ses på figur 6.1
(jævnfør sumkurven i eksempel 5.5)

Ved ***p* - fraktilen** eller $100 \cdot p\%$ fraktilen forstås det tal x_p for hvilket $F(x_p) = p$

Medianen m er 50% fraktilen.

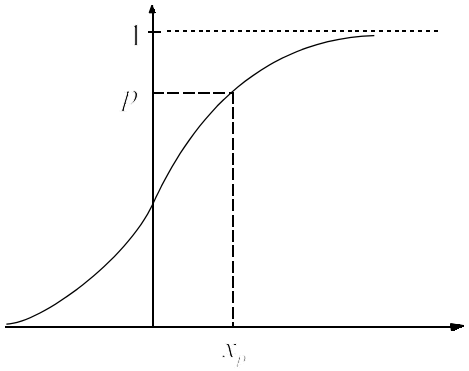


Fig 6.1 *Fordelingsfunktion*

6.4 Beregning af sandsynligheder

Vi har i TI 89: CATALOG\F3\

$p = F(x_p) = P(X \le x_p) = normCdf(-\infty, x_p, \mu, \sigma)$ (**norm**al **C**umulative **d**istribution **f**unktion)

$x_p = F^{-1}(p) = invnormCdf(p, \mu, \sigma)$

Følgende eksempel illustrerer hvordan man på TI 89 beregner sandsynligheder i normalfordelingen.

Eksempel 6.2 Beregning af sandsynligheder i normalfordelingen

Den i eksempel 5.4 angivne stikprøve på 75 patienter fandt vi med rimelighed var normalfordelt. Man finder, at gennemsnit af de 75 tal var $\bar{x}$ =7.287 og spredningen s = 0.134355.
Vi antager derfor, at med tilnærmelse er hele populationen af dårlige knæ i Norden normalfordelt $n(\mu, \sigma)$ med en middelværdi $\mu = 7.29$ og en spredning på $\sigma = 0.134$.

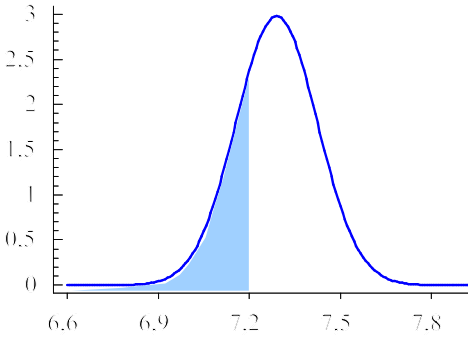
Lad X være pH værdien for et tilfældigt dårligt knæ

1) Beregn sandsynligheden for, at X er mindre end 7.2.
2) Beregn sandsynligheden for at X ligger mellem 7.2 og 7.5
3) Beregn sandsynligheden for, at X er større end eller lig 7.4.
4) Beregn medianen m, dvs. find den værdi m, for hvilke det gælder at $P(X \le m) = 0.50$
5) Beregn 95% fraktilen, dvs. den værdi $x_{0.95}$, for hvilke det gælder at $P(X \le x_{0.95}) = 0.95$

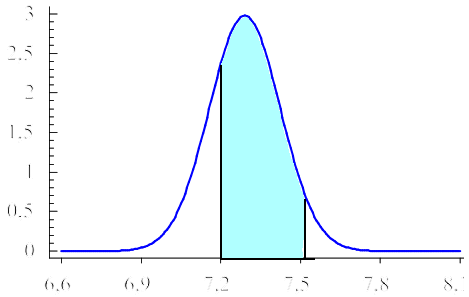
Løsning:

1) Sandsynligheden for, at pH er mindre end 7.2, er lig med arealet af det farvede område under tæthedsfunktionen for n (7.29, 0.134) (se figuren).

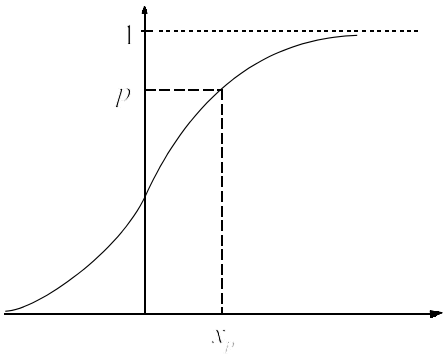
   TI 89: CATALOG\F3\normCdf($-\infty$,7.2,7.29,0.134)
   Resultat: $P(X \le 7.2) = \underline{0.2509}$



2) Ønsker vi tilsvarende at beregne sandsynligheden for, at pH ligger mellem 7.2 og 7.5 er sandsynligheden lig med arealet af det farvede område under kurven på omstående figur.
TI 89: CATALOG\F3\normCdf(7.2,7.5,7.29,0.134)
Resultat $P(7.2 \le X \le 7.5) = \underline{0.629}$



3) Tilsvarende beregnes $P(X \ge 7.4) = \text{normCdf}(7.4, \infty, 7.29, 0.134) = \underline{0.2059}$
4) m = invnorm(0.5,7.29,0.134) = $\underline{7.29}$
5) $x_{0.95}$ = invnorm (0.95, 7.29, 0.134) = $\underline{7.51}$



6.5. Den normerede normalfordeling

Den normerede normalfordeling er bestemt ved at have middelværdien 0 og spredningen 1. En statistisk variabel, der er normalfordelt n (0,1), kaldes sædvanligvis U og dens fordeling ***U*-fordelingen**[1] .
Dens tæthedsfunktion benævnes φ og dens fordelingsfunktion Φ .[2]
Har man ikke et hjælpemiddel til rådighed der som TI 89 kan beregne sandsynligheder i normalfordelingen benytter man en tabel over den normerede normalfordeling. Ud fra denne kan man så beregne sandsynligheder i en vilkårlig normalfordeling.
Dette er unødvendigt når man har et passende hjælpemiddel til rådighed.

[1] I angelsaksiske lande ofte Z og Z-fordelingen.

[2] $\varphi(u) = \frac{1}{\sqrt{2 \cdot \pi}} e^{-\frac{u^2}{2}}$ for ethvert u,og fordelingsfunktionen er bestemt ved $\Phi(u) = P(U \le u) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot \int_{\infty}^{u} e^{-\frac{t^2}{2}} dt$

Imidlertid får man ofte brug for følgende formel, der etablerer en forbindelse mellem de centrale størrelser i en normalfordeling.

$$\boxed{x_p = \mu + u_p \cdot \sigma}$$

Bevis: Lad X være normalfordelt med middelværdi μ og spredning σ.

$U = \dfrac{X-\mu}{\sigma}$ er så også normalfordelt (bevises ikke her)

$$E(U) = E\left(\frac{X-\mu}{\sigma}\right) = \int_{-\infty}^{\infty} \frac{x-\mu}{\sigma} f(x)dx = \frac{1}{\sigma}\int_{-\infty}^{\infty} xf(x)dx - \frac{\mu}{\sigma}\int_{-\infty}^{\infty} f(x)dx = \frac{1}{\sigma}E(x) - \frac{\mu}{\sigma} = 0$$

$$V(U) = V\left(\frac{X-\mu}{\sigma}\right) = \int_{-\infty}^{\infty}\left(\frac{x-\mu}{\sigma}\right)^2 f(x)dx = \frac{1}{\sigma^2}\int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx)\frac{V(X)}{\sigma^2} = 1$$

Heraf ses, at U har middelværdi 0 og spredning 1, dvs. er normeret normalfordelt.

Da $P(X \le x_p) = P\left(\dfrac{X-\mu}{\sigma} \le \dfrac{x_p-\mu}{\sigma}\right) = P\left(U \le \dfrac{x_p-\mu}{\sigma}\right) = \Phi\left(\dfrac{x_p-\mu}{\sigma}\right)$

fås $$P(X \le x_p) = p \Leftrightarrow \Phi\left(\frac{x_p-\mu}{\sigma}\right) = p \Leftrightarrow \Phi^{-1}\left(\Phi\left(\frac{x_p-\mu}{\sigma}\right)\right) = u_p \Leftrightarrow \frac{x_p-\mu}{\sigma} = u_p \Leftrightarrow x_p = \mu + u_p \cdot \sigma$$

◆

Vi vil i det følgende se, hvorledes denne relation med fordel kan anvendes.

Eksempel 6.3. Normalfordeling.

En fabrik støber plastikkasser. Fabrikken får en ordre på kasser, som blandt andet har den specifikation, at kasserne skal have en længde på 90 cm. Kasser, hvis længder ikke ligger mellem 89.2 og 90.8 cm bliver kasseret.

Det vides, at fabrikken producerer kasserne med en længde X, som er normalfordelt med en spredning på 0.5 cm.

1) Hvis X har en middelværdi på 89.6, hvad er så sandsynligheden for, at en kasse har en længde, der ligger indenfor specifikationsgrænserne.
2) Hvor stor er sandsynligheden for at en kasse bliver kasseret, hvis man justerer støbningen, så middelværdien bliver den der giver den mindste procentdel kasserede (spredningen kan man ikke ændre).

Fabrikanten finder, at selv efter den i spørgsmål 2 foretagne justering kasseres for stor en procentdel af kasserne. Der ønskes højst 5% af kasserne kasseret.

3) Hvad skal spredningen σ formindskes til, for at dette er opfyldt?
4) Hvis det er umuligt at ændre σ, kan man prøve at få ændret specifikationsgrænserne. Find de nye specifikationsgrænser (placeret symmetrisk omkring middelværdien 90,0) idet spredningen stadig er 0.5, og højst 5% må kasseres.

En ny maskine indkøbes, og som et led i en undersøgelse af, om der dermed er sket ændringer i middelværdi og spredning produceres 12 kasser ved anvendelse af denne maskine.

Man fandt følgende længder: 89.2 90.2 89.4 90.0 90.3 89.7 89.6 89.9 90.5 90.3 89.9 90.6.

5) Angiv på dette grundlag et estimat for middelværdi og spredning.

Løsning:

1) $P(89.2 < X \le 90.8) =$ normCdf(89.2,90.8,89.6,0.5) = 0.7799 = 78%

2) Middelværdien må nu sættes til midtpunktet af intervallet, dvs. til 90 cm.
$P(X > 90.8) + P(X < 89.2) = 1 - P(89.2 \le X \le 90.8) = 1 -$ normdf(89.2,90.8,90,0.5)=0.1096

3) $P(89.2 < X < 90.8) = 0.95 \Leftrightarrow P(X \le 89.2) = 0.025$

(da der ligger 5% udenfor intervallet, og af symmetrigrunde må så 2,5% ligge på hver sin side af intervallet.)

Metode 1:Ved indsættelse i ligningen $x_{0.025} = \mu + u_{0.025} \cdot \sigma$ fås nu

$$89.2 = 90 + u_{0.025} \cdot \sigma \Leftrightarrow \sigma = \frac{89.2 - 90}{u_{0.025}} \Leftrightarrow \sigma = \frac{-0.08}{\text{invnormCdf}(0.025,0,1)} \Leftrightarrow \sigma = 0.4082$$

Metode 2: Ligningen $P(X \le 89.2) = 0.025$ løses med hensyn til σ

solve(normCdf$(-\infty, ,89.2,90,x) = 0.025, x) | x > 0$ Resultat: $x = 0.4082$

4) Af symmetrigrunde (samme begrundelse som under punkt 3) fås:
$P(90.0 - d < X < 90.0 + d) = 0.95 \Leftrightarrow P(X \le 90.0 - d) = 0.025$ og $P(X \le 90.0 + d) = 0.975$.
Vi får nedre grænse =invnormCdf(0,025,90,0,5) = 89,02 = 89.0
Øvre grænse =invnormCdf(0,975;90;0,5) = 90,98 = 91.0

5) APPS\Stats/List . Tallene indtastes i list1
F4\1:1-var Stats
I den fremkomne menu sættes "list" til "List1" (benyt eventuelt Var-Link til at finde List1)
I udskriften findes $\bar{x} = 89.97$ og $s = 0.435$ ◆

6.6 Konfidensinterval for middelværdi

6.6.1. Indledning

Udtages en stikprøve fra en population er det jo for, at man ud fra stikprøven kan fortælle noget centralt om hele populationen.
For en normalfordelt variabel X har vi således som estimat (skøn) for populationens middelværdi μ sat stikprøvens gennemsnit $\bar{x}$, og som et estimat for populationens spredningen (standardafvigelse) σ sat stikprøvens spredning s.
Et gennemsnit er jo altid behæftet med en vis usikkerhed.
Det er derfor ikke nok, at angive at den "sande" middelværdi er $\bar{x}$, vi må også angive et "usikkerhedsinterval".
Et interval indenfor hvilket den sande værdi" μ med eksempelvis 95% sikkerhed vil ligge kaldes et **95% konfidensinterval**.

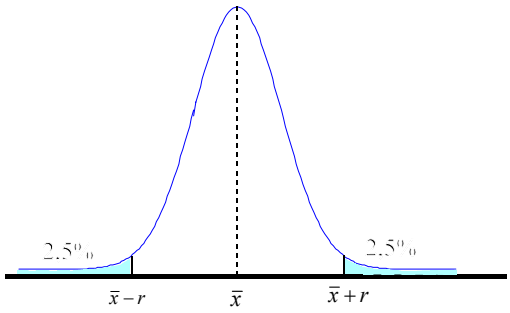
Sætning 6.1: Gennemsnits spredning og fordeling
Lad $\bar{x}$ være gennemsnittet af værdierne i en stikprøve på *n* tal.
1) $\bar{x}$ vil være tilnærmelsesvis normalfordelt, hvis blot *n* er tilstrækkelig stor (*i praksis over 30*).
2) Spredningen på $\bar{x}$ er $\sigma(\bar{x}) = \frac{\sigma}{\sqrt{n}}$, hvor σ er spredningen på den enkelte værdi i stikprøven.

Punkt 1 sikrer, at selv om værdierne i stikprøven ikke er normalfordelt, så vil gennemsnittet være det, blot n er stor (over 30)
Punkt 2 viser, at gennemsnittet kan man "stole" mere på end den enkelte måling, da den har en mindre spredning.

Eksempel 6.4 Beregning af spredning af gennemsnit
Find spredningen på gennemsnittet for følgende 12 målinger
89.2 90.2 89.4 90.0 90.3 89.7 89.6 89.9 90.5 90.3 89.9 90.6.
Løsning:
I eksempel 6.3 fandt vi at de 12 målinger
havde $\bar{x} = 89.97$ og s = 0.435
Heraf følger, at spredningen på gennemsnittet er $\frac{s}{\sqrt{n}} = \frac{0.435}{\sqrt{12}} = \underline{\underline{0.126}}$

6.6.2 Beregning af konfidensinterval, hvis spredning er kendt eksakt.

Et 95% konfidensinterval $[\bar{x} - r; \bar{x} + r]$ må ligge symmetrisk omkring gennemsnittet, og således, at $P(\bar{x} - r \le \bar{X} \le \bar{x} + r) = 0.95$.
Heraf følger, at hvis den sande middelværdi μ ligger i et af de skraverede områder, så er der mindre end 2.5% chance for, at vi ville have fået det fundne gennemsnit $\bar{x}$.
For at finde grænsen for intervallet, må vi finde en middelværdi μ så $P(\bar{X} \le \bar{x}) = 0.025$.
Lad os illustrere det ved følgende eksempel:



Eksempel 6.5 Beregning af 95% konfidensinterval
Lad gennemsnittet af 12 målinger være $\bar{x} = 90$
Lad os antages at spredningen kendes eksakt til $\sigma = 0.5$
Hvis den sande middelværdi μ afviger stærkt fra 90 er det yderst usandsynligt, at vi ville have fået et gennemsnittet på 90.
Eksempelvis, hvis $\mu = 92$ bliver $P(\bar{X} \le 90) = \text{normCdf}(-\infty, 90, 92, 0.5/\sqrt{\ }(12)) = 0$
dvs. det er ganske usandsynligt at den sande middelværdi var 92.
For at finde grænsen kunne man finde μ af ligningen

$P(\overline{X} \le 90) = 0 \Leftrightarrow$ solve(normCdf($-\infty$,90,x,0.5/$\sqrt{\ }$(12))=0.025,x) Resultat x = 90.283
95% konfidensinterval [90 -0.283;90 +0.283] = [89.717 ; 90.283]

Lettere er det at benytte formlen $x_p = \mu + u_p \cdot \sigma$ som ved benyttelse af, at $\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}$ giver $\mu = \bar{x} - u_{0.025} \cdot \frac{\sigma}{\sqrt{12}}$. Indsættes $u_{0.025} = invnorm(0.025,0,1) = -1.96$ fås, at øvre grænse for konfidensintervallet er $\mu = \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{12}} = 90 + 1.96 \cdot \frac{0.5}{\sqrt{12}} = 90.283$.
Da der er symmetri omkring $\bar{x}$ fås samme konfidensinterval som før ◆

Som det fremgår af eksempel 6.5 gælder følgende

Er spredningen eksakt kendt er et 95% konfidensinterval bestemt ved formlen

$$\bar{x} - u_{0.975} \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + u_{0.975} \cdot \frac{\sigma}{\sqrt{n}} \qquad (1)$$

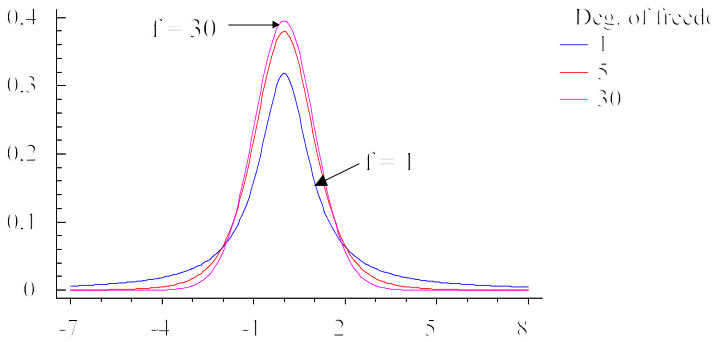
Ønskes eksempelvis et 99% eksakt skal $u_{0.975}$ erstattes med $u_{0.995}$ osv.

6.6.3. Beregning af konfidensinterval hvis spredning ikke kendt eksakt

Sædvanligvis er populationens spredning σ jo ikke eksakt kendt, men man regner et estimat *s* ud for den.
Da *s* jo også varierer fra stikprøve til stikprøve, giver dette en ekstra usikkerhed, så konfidensintervallet for μ bliver bredere.
Hvis stikprøvestørrelsen er over 30 er denne usikkerhed dog uden væsentlig betydning, så i sådanne tilfælde kan man i formel (1) blot erstatte σ med *s*.

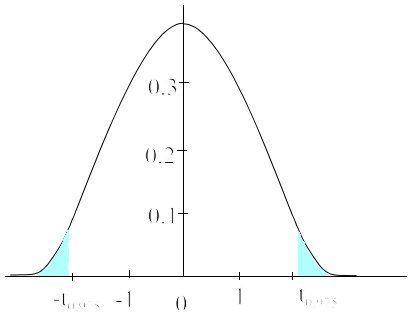
Er stikprøvestørrelsen under 30 bliver denne usikkerhed på s så stor, at man i formel (1) må erstatte U- fraktilen $u_{0.975}$ med en såkaldt t - fraktil $t_{0.975,f}$.

t-fordelinger
En t - fordeling har samme klokkeformede udseende som en U - fordeling, men i modsætning til U - fordelingen afhænger dens udseende af antallet *n* af tal i stikprøven. Er $f = n$ -1 stort (over 30) er forskellen mellem en U- fordeling og en t- fordeling meget lille. Er *f* lille bliver t - fordelingen bredere end U - fordelingen.
Tallet $f = n$ - 1 kaldes **frihedsgradstallet.**
Grafen nedenfor viser tæthedsfunktionen for *t*-fordelingerne for f = 1, 5 og 30.



Ved t - fraktilen $t_{0.975,12}$ forstås 0.975 - fraktilen med frihedsgradstallet 12.

Eksempel 6.6. Beregning af t-fraktiler.
Find fraktilerne $t_{0.975,12}$ og $t_{0.025,12}$.
Løsning:
Af symmetrigrunde (se figuren) er de 2 fraktiler lige store med modsat fortegn, dvs.
$t_{0.025,12} = -t_{0.975,12}$
TI 89: CATALOG\F3\inv_t(0.975,12) Resultat: 2.1788



Et 95 % konfidensinterval er bestemt ved formlen:

$$\bar{x} - t_{0,975,n-1} \cdot \frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{0.975,n-1} \cdot \frac{s}{\sqrt{n}} \qquad (2)$$

Eksempel 6.6. Konfidensinterval, hvis spredningen ikke er kendt eksakt.
En forstmand er interesseret i at bestemme middelværdien af diameteren af voksne egetræer i en bestemt fredet skov.
Der blev målt diameteren på 7 tilfældigt udvalgte egetræer (i 1 meters højde over jorden)
Resultatet ses i følgende skema.

diameter (cm)	64.0	33.4	45.8	56.0	51.5	29.2	63.7

1) Beregn $\bar{x}$ og s.
2) Beregn et 95% konfidensinterval for middelværdien μ.

Løsning:
APPS\STAT/LIST : Data indtastes. i list1

Metode 1
1) CATALOG\mean(list1) = 49,08571 og stdDev(list1) = 13.7957
2) CATALOG\F3\inv_t(0.975,6) = 2.4469
   Man indsætter nu de fundne tal som jo alle står i "home" i formlen
   Øvre grænse: 49.085+2.57*13.7957/$\sqrt{\ }$ (7) = 61.84
   Nedre grænse: 49.085-2.57*13.7957/$\sqrt{\ }$ (6) = 36.33
   95% konfidensinterval: [36.33 ; 61.84]

Metode 2
1) F4\1-Var Stats\ENTER Blandt udskrift findes $\bar{x}$ = 49,08571 s = 13,7957
2) APPS\STAT/LIST \F7\ T Interval
   I Menu: Data Input Method sættes til "Data" \ENTER\ENTER
   Resultat: CInt = [36.33 ; 61.84]

Opgaver

Opgave 6.1

1) En stokastisk variabel X er normalfordelt med $\mu = 0$ og $\sigma = 1$.
   Find $P(X \le 0.75)$, $P(X > 1.6)$ og $P(0.75 < X < 1.6)$.
2) En stokastisk variabel X er normalfordelt med $\mu = 25.1$ og $\sigma = 2.4$.
   Find $P(22.3 < X \le 27.8)$.

Opgave 6.2

En "soft-drink" maskine er reguleret, så den i middel fylder 200 ml i en kop. Rumfanget X antages at være normalfordelt med en spredning på 15 ml.

1) Hvor stor en brøkdel af kopperne vil indeholde mindre end 224 ml
2) Hvad er sandsynligheden for at en kop indeholder mellem 191 og 209 ml
3) Kopperne kan højst indeholde 230 ml.
   Hvor mange af de næste 1000 fyldte kopper, vil i middel være overfyldt.
4) 25% af kopperne vil i middel indeholde mindre end $x_{0.25}$ ml. Find $x_{0.25}$.
5) 95% af kopperne vil indeholde mere end a ml. Find a

Opgave 6.3

Maksimumstemperaturen, der opnås ved en bestemt opvarmningsproces, har en statistisk fordeling med en middelværdi på 113.3° og en spredning på 5.6°C. Det antages, at maksimumstemperaturens variation er tilfældig og kan beskrives ved en normalfordeling.

1) Find procenten af maksimumstemperaturer, der er mindre end 116.1°C.
2) Find procenten af maksimumstemperaturer, der ligger mellem 115°C og 116.7°C.
3) Find den værdi, som overskrides af 57.8% af maksimumstemperaturerne.

Man overvejer at gå over til en anden opvarmningsproces. Man udfører derfor 16 gange i løbet af en periode forsøg, hvor man måler maksimumstemperaturen, der opnås ved denne nye proces. Resultaterne var 116.6 , 116,6 , 117,0 , 124,5 , 122,2 , 128,6 , 109,9 , 114,8 , 106,4 , 110,7, 110,7 , 113,7 , 128,1, 118,8 , 115,4 , 123,1

4) Giv et estimat for middelværdien og spredningen.

Opgave 6.4

En topedo affyres mod et 250 meter bredt mål.
Man sigter efter målets midtpunkt. Afstanden fra midtpunktet til det punkt der rammes er normalfordelt med en middelværdi på 0 og en spredning σ på 100 meter.
Beregn sandsynligheden for at man rammer målet.

Opgave 6.5

En automatisk dåsepåfyldningsmaskine fylder hønskødssuppe i dåser. Rumfanget er normal-fordelt med en middelværdi på 800 ml og en spredning på 6,4 ml.

1) Hvad er sandsynligheden for, at en dåse indeholder mindre end 790 ml?.
2) Hvis alle dåser, som indeholder mindre end 790 ml og mere end 805 ml bliver kasseret, hvor stor en procentdel af dåserne bliver så kasseret?
3) Bestem de specifikationsgrænser der ligger symmetrisk omkring middelværdien på 800 ml, og som indeholde 99% af alle dåser.

Opgave 6.6
En fabrik planlægger at starte en produktion af rør, hvis diametre skal opfylde specifikationerne 2,500 cm ± 0,015 cm.
Ud fra erfaringer med tilsvarende produktioner vides, at de producerede rør vil have diametre, der er normalfordelte med en middelværdi på 2,500 cm og en spredning på 0,010 cm. Man ønsker i forbindelse med planlægningen svar på følgende spørgsmål:
1) Hvor stor en del af produktionen holder sig indenfor specifikationsgrænserne.
2) Hvor meget skal spredningen σ ned på, for, at 95% af produktionen holder sig indenfor specifikationsgrænserne (middelværdien er uændret på 2,500 cm).
3) Fabrikken overvejer, om det er muligt at få indført nogle specifikationsgrænser (symmetrisk omkring 2,500), som bevirker, at 95% af dets produktion falder indenfor grænserne. Find disse grænser, idet det stadig antages at middelværdien er 2.500 og spredningen 0.010 cm.

Opgave 6.7
Trykstyrken i beton blev kontrolleret ved at man støbte 12 betonklodser og testede dem. Resultatet var:

2216	2225	2318	2237	2301	2255	2249	2281	2275	2204	2263	2295

1) Find et estimat for trykstyrkens middelværdi μ og spredning σ.
2) Angiv et 95% konfidensinterval for μ.

Opgave 6.8
En fabrik producerer stempelringe til en bilmotor. Det vides, at stempelringenes diameter er approksimativt normalfordelt. Stempelringene bør have en diameter på 74.036 mm og en spredning på 0.001 mm. For at kontrollere dette udtog man tilfældigt 15 stempelringe af produktionen og målte diameteren. I resultaterne har man for simpelheds skyld, kun angivet de 3 sidste cifre, altså 74.0365 angives som 365. Man fandt følgende resultater

342	364	370	361	351	368	357	374	340	362	378	384	354	356	369

1) Find et estimat for ringenes diameter μ og spredning σ.
2) Angiv et 99% konfidensinterval for μ.
3) Angiv et 99% konfidensinterval for μ, når man fra tidligere målinger ved, at $\sigma = 0.001$.

Opgave 6.9
Ved en fabrikation af et bestemt sprængstof er det vigtigt, at en reaktoropløsning har en pH-værdi omkring 8.50. Der foretages 6 målinger på en bestemt reaktantopløsning. Resultaterne var:

pH	8.54	7.89	8.50	8.21	8.15	8.32

Den benyttede pH-målemetode antages på baggrund af tidligere lignende målinger at give normalfordelte resultater.
1) Angiv et estimat for opløsningens middelværdi og spredning.
2) Angiv et 95% konfidensinterval for pH.

Opgave 6.10
De 10 øverste ark papir i en pakke med printerpapir har følgende vægt

4.21	4.33	4.26	4.27	4.19	4.30	4.24	4.24	4.28	4.24

Angiv 95%-konfidensintervaller for middelværdi og spredning af papirets vægt.

Opgave 6.11
Til undersøgelse af alkoholprocenten i en persons blod foretages 4 uafhængige målinger, som gav følgende resultater (i ‰):

108	102	107	98

1) Opstil et 95% konfidensinterval for personens alkoholkoncentration.

Facitliste for udvalgte opgaver

Kapitel 1

1.1 0.1 0.5 0.8 0.2 0.7
1.2 (1) 0.9134 (2) 0.9678
1.3 (1) 8.75% (2) 38.75% (3) 41.25% (4)11.25%
1.4 (1) 6.4% (2) 78.4% (3) 7.2%
1.5 (1) 27.1% 36.0% 9.756% (2) 53.34% (3) 49.20%
1.6 (1) 41.67% (2) 12%

Kapitel 2

2.1 (a) 6 (b) 24
2.2 (1) 100 (2) 2400
2.3 3^{40}
2.4 31
2.5 30.24%
2.6
2.7 $1.283 \cdot 10^{12}$
2.8 (a) - (b) 736
2.9 30
2.10
2.11 (A) 0.018% (B) 1.29% (C) 38.24%
2.12 (1) 0.435% (2) 49.57% (3) 41.30%
2.13 (1) 91.67% (2) 25.00% (3) 9.167%
2.14 (1) 17.68% (2) 59.28%
2.15

Kapitel 3

3.1
3.2
3.3
3.4
3.5
3.6

Kapitel 4

4.1 (a) 27.87% (b) 68.46%
4.2 (a) 34.10% (b) 40
4.3 37.11%
4.4 0.0275%
4.5 (a) 9.05% (b) 25
4.6 19.77%
4.7
4.8 94.5%
4.9

4.10 77.86%
4.11 5.83%
4.12 12.85%
4.13 (1) 7.94% (2) 11.8%

4.15 (1) 0.108 (2) [0.089 ; 0.127] (3) 21.04
4.16 (1) [0.30 ; 0.38] (b) 784
4.18 [0.03 ; 0.07]
4.19 (1) [0.04;0.10] (2) ca 625
4.20 (1) 0.683 (2) [0.600 ; 0.767] (3) 322

Kapitel 5

5.1 (1) - (2) ca 24%
5.2 (1) - (2) ca 0.052

Kapitel 6

6.1 (1) 0.7734 0.0548 0.1718 (2) 0.7480
6.2
6.3 (1) 69.15% (2) 10.88% (3) 114.4 (4) 117.3 6.535
6.4 78.87%
6.5 (1) 5.91% (2) 27.64% (3) [783.51; 816.49]
6.6 (1) 86.64% (2) 0.008 (3) 2.48 2.52
6.7 (1) 2259.92 35.569 (2) [2237 ; 2283] (3) 25
6.8 (1) 74.0362 0.00124 (2) [74.035; 74.037] (3) 43
6.9 (1) 8.268 0.241 (2) [8.02 ; 8.52] (3) 27
6.10 [4.23 ; 4.29]
6.11 0.965 ; 1.111]

STIKORD