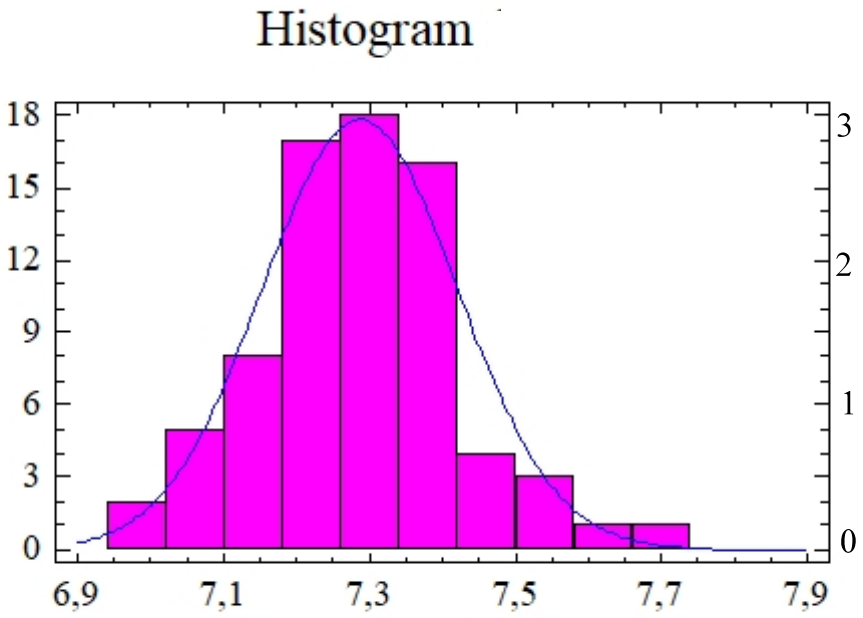
MOGENS ODDERSHEDE LARSEN

STATISTISKE GRUNDBEGREBER



17 udgave 2016

FORORD

Der er i denne bog søgt at give letlæst og anskuelig fremstilling af de statistiske grundbegreber til brug ved en indledende undervisning i statistik. De væsentligste definitioner og sætninger forklares derfor fortrinsvist ved hjælp af figurer og gennemregnede praktiske eksempler. Ønskes en mere matematisk uddybende forklaring, bevis for sætninger osv. kan dette ofte findes i et særskilt tillæg til bogen, som findes på nettet under titlen "Supplement til statistiske grundbegreber".

Læsning: Bogen er bygget således op, at der hurtigt nås frem til normalfordelingen og de vigtige normalfordelingstest. Disse vigtige begreber kan derfor blive grundigt indarbejdet, selv om der kun er kort tid til rådighed. Er det af tidsmæssige grunde svært at nå hele bogen kan man uden skade for helheden overspringe kapitel 9, ligesom man eventuelt kan tage kapitlerne 1 og 8 mere oversigtsagtigt.
Sidst i hver kapitel findes en række opgaver, der yderligere kan fremme forståelsen.
Bagerst i bogen findes en facitliste til alle opgaverne.

I et længere kursusforløb er denne bog tænkt at skulle efterfølges af M. Oddershede Larsen: Videregående Statistik", som kan hentes gratis på e-mailadressen www.larsen-net.dk

Regnemidler. Det er hensigtsmæssigt, at man har adgang til et program med de sædvanlige statistiske fordelinger indbygget.
I eksemplerne angives således, hvorledes beregningerne kan foretages med det meget udbredte regneark Excel.
Ønskes i stedet at anvende det TI-Nspire (PC-udgaven) , kan man på hjemmesiden www.larsen-net.dk under statistik 2 hente den samme bog

I 8- udgave findes tabeller over de sædvanlige statistiske funktioner, samt forklaret hvordan tabellerne anvendes

Denne udgave, samt 8 udgave kan sammen med en række andre noter findes på adressen:
www.larsen-net.dk

31. august 2017 Mogens Oddershede Larsen

INDHOLD

1 INTRODUKTION TIL STATISTIK

Ved næsten alle ingeniørmæssige problemer vil de indsamlede data udvise **variation.** Måler man således gentagne gange indholdet (i %) af et bestemt stof i et levnedsmiddel, vil det procentvise indhold ikke blive præcis samme tal for hver gang man foretager en måling. Dette kunne naturligvis være en usikkerhed ved målemetoden, men det vil sjældent være den væsentligste årsag.

Ved mange industrielle processer vil en række ukontrollable forhold indvirke på det endelige resultat. Eksempelvis vil udbyttet af en kemisk proces variere fra dag til dag, fordi man ikke har fuldstændig kontrol over **forsøgsbetingelser** som temperatur, omrøringstid, tidspunkt for tilsætning af råmaterialer, fugtighed osv. Endvidere er **forsøgsmaterialerne** muligvis ikke homogene nok. Råmaterialerne kan f.eks. være af varierende kvalitet, der må bruges forskelligt apparatur under produktionsprocessen, forskelligt personale deltager i arbejdet osv.

Statistik drejer sig om at samle, præsentere og analysere data med henblik på at foretage beslutninger og løse problemer.

I den **deskriptive statistik** beskrives data ved tabeller, grafisk (lagkagediagrammer, søjlediagrammer) og ved beregning af karakteristiske tal såsom gennemsnit og spredning.
Man kan eksempelvis i "Danmarks Statistik" (findes på nettet under adressen www.statistikbanken.dk) finde, hvor mange personbiler der er i Danmark i 2009 opdelt efter alder.
Man kender her **populationen** (biler i Danmark), kan grafisk vise deres fordeling i et søjlediagram og beregne deres gennemsnitlige alder.

I den mere analyserende statistik (kaldet **inferentiel statistik**) søger man ved mere avancerede statistiske metoder ud fra en repræsentativ stikprøve at konkludere noget om hele populationen.

Eksempelvis udtages ved en meningsmåling en forhåbentlig repræsentativ **stikprøve** på 1000 vælgere, som man spørger om hvilket politisk parti de ville stemme på, hvis der var valg i morgen.
Man vil så ud fra stikprøven konkludere, at hvis man spurgte hele populationen (alle vælgere i Danmark) , så ville man med en vis usikkerhed få samme resultat.
Viser stikprøven, at partiet "Venstre" vil gå 2.5% tilbage, så vil det samme ske, hvis der var valg i morgen.
Et sådant tal er naturligvis usikkert. Man må derfor anvende passende statistiske metoder til eksempelvis at beregne, at usikkerheden er på 2%.

2. DESKRIPTIV STATISTIK

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").

2.1 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.

De følgende 2 eksempler viser anvendelse af henholdsvis lagkagediagram og søjlediagram

Eksempel 2.1 Lagkagediagram

Nedenfor er angivet hvordan en kommunes udgifter fordeler sig på de forskellige områder.

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

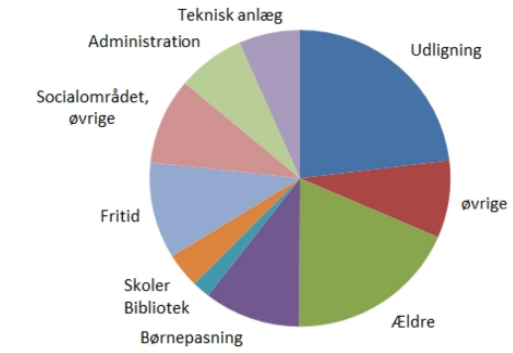
Dan et lagkagediagram til anskueliggørelse heraf.

Løsning:

Husk først at indlægge tilføjelsesprogrammer (se kapitel 10)

Løsning:

Data indsættes i 2 kolonner.
Marker de 2 kolonner ▶ Vælg på værktøjslinien "Indsæt" ▶ Cirkel ▶ Cursor på figur ▶ Formater dataetiketter ▶
Vælg "kategorinavn" og "udenfor".



Eksempel 2.2 Søjlediagram

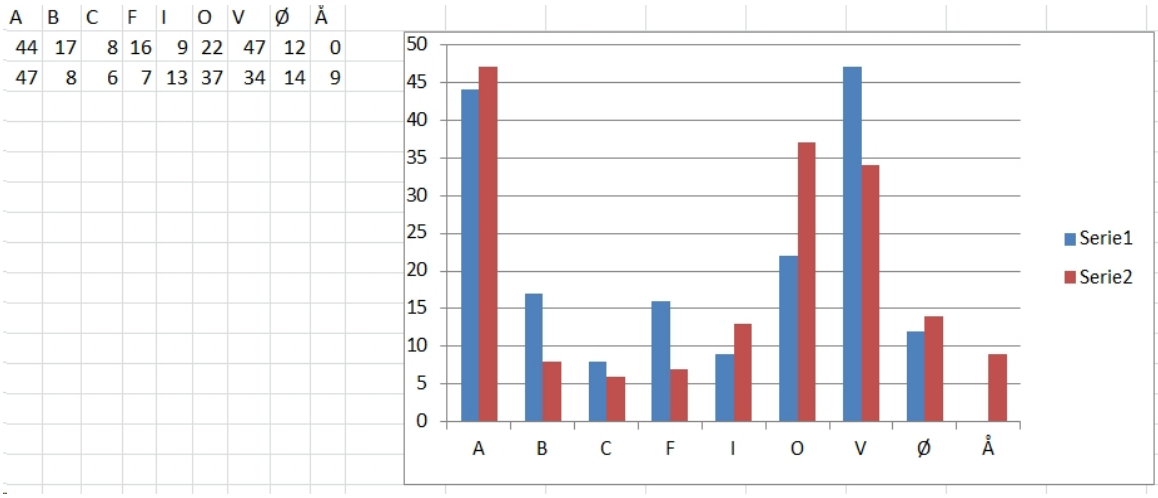
Følgende tabel angiver mandattallet ved to folketingsvalg.

Partier		A	B	C	F	I	O	V	Ø	Å
Mandater	2015	47	8	6	7	13	37	34	14	9
	2011	44	17	8	16	9	22	47	12	0

A=Socialdemokraterne, B=Radikale venstre, C=Konservative folkeparti, F=Socialistisk folkeparti, I=Liberal alliance, O=Dansk Folkeparti, V=Venstre, Ø=Enhedslisten, Å=Alternativet
Anskueliggør disse mandattal ved at tegne et søjlediagram

Løsning:

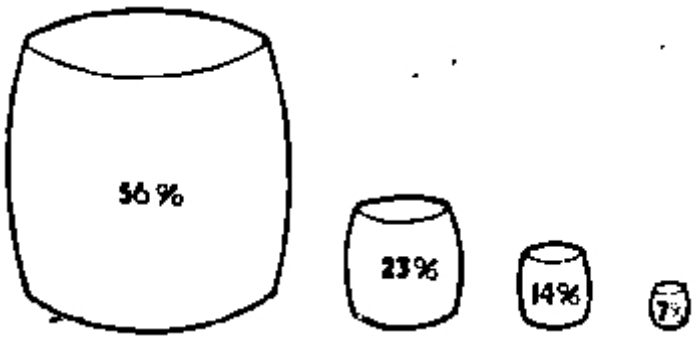
Som i eksempel 2.1 blot vælges ▶Søjle



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 2.3. 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.



De mest almindelige figurer til at give et visuelt overblik over større talmaterialer er histogrammer (søjlediagrammer) og kurver i et koordinatsystem.

2.2. KVANTITATIVE DATA (VARIABLE)

Kvantitative data er data, hvor registreringen i sig selv er tal, der angiver en bestemt rækkefølge, f. eks. som i eksempel 2.4 hvor data registreres efter det tidspunkt hvor registreringen foregår eller som i eksempel 2.5, hvor det er størrelsen af registrerede værdi der er af interesse.

Eksempel 2.4. Kvantitativ variabel: tid

Fra "statistikbanken (adresse **http://www.statistikbanken.dk/)** er hentet følgende data ind i Excel, der beskriver hvorledes indvandringer og udvandringer er sket gennem tiden.

Excel: Vælg "Befolkning og valg" ▶ Flytning til og fra udlandet ▶ Ind- og udvandring på måneder ▶ under "bevægelse" vælges "flere valgmuligheder", marker alle ▶ under "måned" vælges "flere valgmuligheder" år og derefter alle ▶ Tryk på tabel ▶ Drej tabel med uret ▶ Gem som Excel fil

Indvandringer og udvandringer efter tid og bevægelse

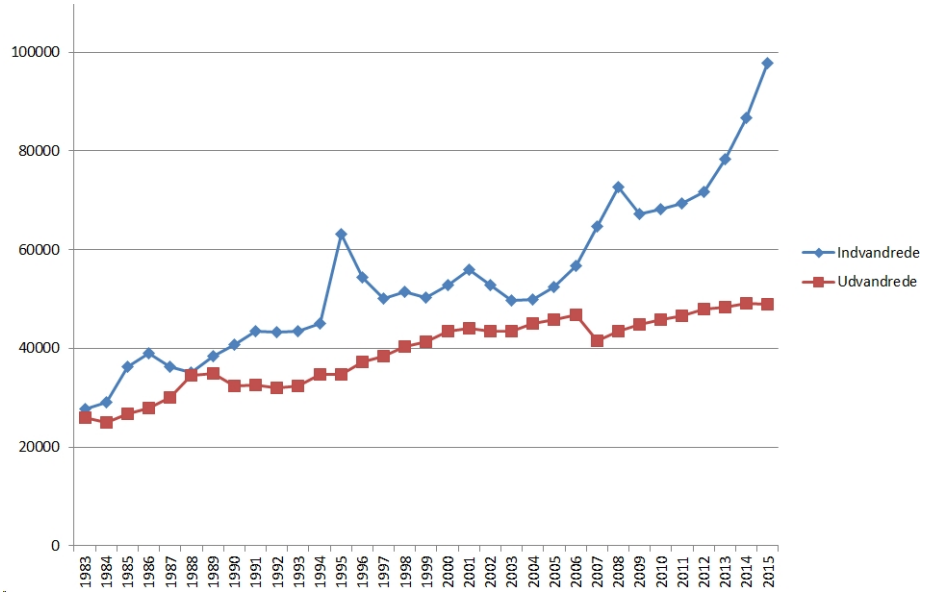
	Indvandrede	Udvandrede
1983	27718	25999
1984	29035	25053
1985	36214	26715
1986	38932	27928
1987	36296	30123
1988	35051	34544
1989	38391	34949
1990	40715	32383
1991	43567	32629
1992	43377	31915
1993	43400	32344
1994	44961	34710
1995	63187	34630
1996	54445	37312
1997	50105	38393
1998	51372	40340
1999	50236	41340
2000	52915	43417
2001	55984	43980
2002	52778	43481
2003	49754	43466
2004	49860	45017
2005	52458	45869
2006	56750	46786
2007	64656	41566
2008	72749	43490
2009	67161	44874
2010	68282	45882
2011	69298	46684
2012	71739	49988

Giv en grafisk beskrivelse af disse data.

Løsning

Da dataene er registreret efter tid (år) (den kvantitative variabel "tid") tegnes to kurver i samme koordinatsystem:

Marker udskriftsområde (data) ▶ Vælg på værktøjslinien " Indsæt" ▶ Streg ▶ Marker ønsket figur



Eksempel 2.5. Kvantitativ variabel , størrelse af brintionkoncentrationen pH

I menneskers led udskiller den inderste hinde en "ledvæske" som "smører" leddet. For visse ledsygdomme kan brintionkoncentrationen (pH) i denne væske tænkes at have betydning. Som led i en nordisk medicinsk undersøgelse af en bestemt ledsygdom udtog man blandt samtlige patienter der led af denne sygdom en repræsentativ stikprøve ved simpel udvælgelse 75 patienter og målte pH i ledvæsken i knæet.

Resultaterne (som kan findes som excel-fil på adressen www.larsen-net.dk) var følgende:

7.02	7.26	7.31	7.16	7.45	7.32	7.21	7.35	7.25	7.24	7.20	7.21	7.27	7.28	7.19
7.39	7.40	7.33	7.32	7.35	7.34	7.41	7.28	7.27	7.28	7.33	7.20	7.15	7.42	7.35
7.38	7.32	7.71	7.34	7.10	7.35	7.15	7.19	7.44	7.12	7.22	7.12	7.37	7.51	7.19
7.30	7.24	7.36	7.09	7.32	6.95	7.35	7.36	7.52	7.29	7.31	7.35	7.40	7.23	7.16
7.26	7.47	7.61	7.23	7.26	7.37	7.16	7.43	7.08	7.56	7.07	7.08	7.17	7.29	7.20

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.

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 7.71 og mindste tal er 6.95 og variationsbredden derfor 7.71 - 6.95 = 0.76.

Dernæst deles tallene op i et passende antal intervaller (klasser). Som det første bud vælges ofte et antal nær $\sqrt{n}$. Da $\sqrt{75} \approx 9$ vælges ca. 9 klasser. Da $\frac{0.76}{9} \approx 0.08$ deler vi op i de klasser, der ses af tabellen. Dette giver 10 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).

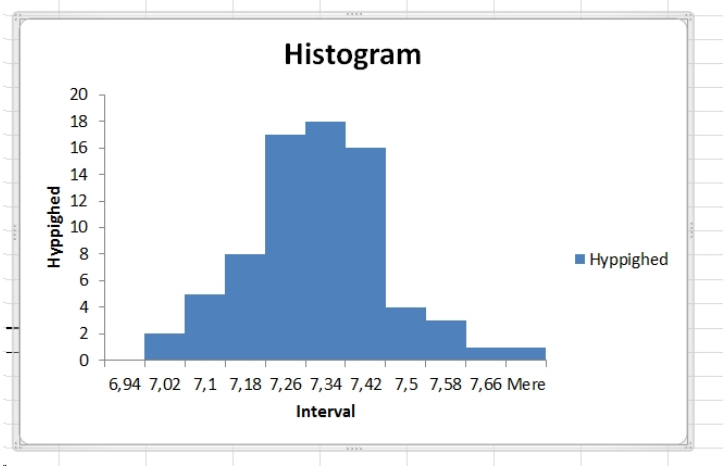
Klasser		Antal *n*
]6.94 - 7.02]	//	2
]7.02 - 7.10]	/////	5
]7.10 - 7.18]	////////	8
]7.18 - 7.26]	//////////////////	17
]7.26 - 7.34]	//////////////////	18
]7.34 - 7.42]	//////////////////	16
]7.42 - 7.50]	////	4
]7.50 - 7.58]	///	3
]7.58 - 7.66]	/	1
]7.66 - 7.74]	/	1

Allerede her kan man se, at antallet er størst omkring 7.30, og så falder hyppigheden nogenlunde symmetrisk til begge sider.

Data indtastes i eksempelvis søjle A1 til A75 (data findes på adressen www.larsen-net.dk)
Vælg "Data"►Dataanalyse ►Histogram
I den fremkomne tabel udfyldes "inputområdet" med A1:A75 og man vælger "diagramoutput"..
Trykkes på OK fås en tabel med hyppigheder, og en figur, hvor intervalgrænserne er fastlagt af Excel.
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 6.94, i B2 7.02 osv. til B10: 7.66) og så skrive B1:B10 i intervalområdet.
Man vælger outputområdet, og markerer Diagramoutput
Da et histogram har søjlerne samlet, foretages følgende: cursor på en søjle ►tryk højre musetast ►formater dataserie ►indstilling ►mellemrumsbredde = 0 ►ok

Der fremkommer følgende histogram (efter at have valgt udskrift med 2 decimaler):

Interval	Hyppighed
6,94	0
7,02	2
7,1	5
7,18	8
7,26	17
7,34	18
7,42	16
7,5	4
7,58	3
7,66	1
Mere	1



Histogrammet er et "klokkeformet histogram", hvor der er flest tal fra 7.19 til 7.42, og derefter falder antallet 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 siger, at resultaterne er normalfordelt (beskrives nærmere i næste kapitel)

2.3 KARAKTERISTISKE TAL

Skal man sammenligne to talmaterialer, eksempelvis sammenligne de 75 pH-værdier i eksempel 1.4 med 200 dårlige knæ fra Tyskland, har det ingen mening at sammenligne hyppighederne

Man må i sådanne tilfælde angive nogle tal, som gør det muligt at foretage en sammenligning. Dette kunne blandt andet ske ved at man udregnede de relative hyppigheder

2.3.1 Relativ hyppighed

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

I eksempel 2.5 er den relative hyppighed for pH - værdier i intervallet]7.18 - 7.26]:

$$\frac{17}{75} = 0.2267 = 22.57\%$$

Man kunne sige, at "sandsynligheden" er 22.57% for at pH ligger i dette interval.

2.3.2 Middelværdi og spredning.

Middelværdi, gennemsnit.

Kendes hele "populationen" (målt højden på **alle** danske mænd) kan beregnes en "korrekt middterværdi" kaldet middelværdi μ (græsk my)

Ud fra stikprøven vil en tilnærmet værdi (kaldet et **estimat**) for μ være **gennemsnittet** $\bar{x}$ (kaldt x streg).

Kaldes observationerne i en stikprøve $x_1, x_2, \ldots, x_n$ er $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$

Eksempel 2.6: Gennemsnit

Find gennemsnittet af tallene 6, 17, 7, 13, 5, 3

Løsning:

Håndregning: $\bar{x} = \frac{6+17+7+13+5+3}{6} = 8.5$

Tast tallene i en kolonne ▶Vælg på værktøjslinien *fx* ▶Statistisk▶ Middel(A1..A6)

A	B	C
6		
17	MIDDEL(A1:A6)	8,5
7		
13		
5		
3		

Spredningsmål

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.

Spredning (engelsk: standard deviation)
Hvis spredningen baserer sig på hele populationen benævnes den σ (sigma) .
Baserer spredningen sig kun på en stikprøve benævnes den s.
Man siger, at s er et estimat (skøn) for σ .

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

Kvadratsummen $\sum_{i=1}^{n}(x_i - \bar{x})^2$ benævnes kort SAK (Summen af Afvigelsernes Kvadrater) eller SS (Sum of Squares)

Ved **variansen** for en stikprøve forstås s^2 .

Eksempel 2.7: Spredning

Find varians og spredning af tallene 6, 17, 7, 13, 5, 3

Løsning:

I eksempel 2.6 findes gennemsnittet $\bar{x} = 8.5$

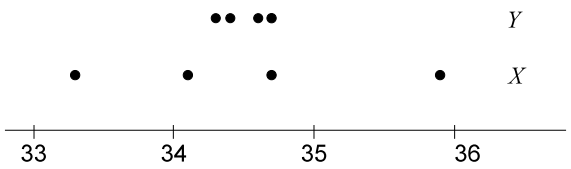
Håndregning: $s^2 = \frac{(6-8.5)^2 + (17-8.5)^2 + (7-8.5)^2 + (13-8.5)^2 + (5-8.5)^2 + (3-8.5)^2}{6-1} = \underline{\underline{28.7}}$

Spredningen $s = \sqrt{28.7} = \underline{\underline{5.357}}$

Excel: Som eksmpel 2.6, men nu vælges varians og stdafv.s

VARIANS.S(A1:A6)	28,7
STDAFV.S(A1:A6)	5,357

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 eksempel 2.7, at der er 5 frihedsgrader, da kendskab til de første 5 led på 6, 17, 7, 13, 5 er nok til at bestemme det sjette 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 såkaldte normalfordelte resultater, er de tilsvarende tal ca. 95% og 99.7 %
I eksempel 2.7 fandt vi således $\bar{x} - 2 \cdot s$ = 8.5-2· 5.357=-2.21 og

$$\bar{x} + 2 \cdot s = 8.5 + 2 \cdot 5.357 = 19.21$$

Det ses. at alle tallene ligger indenfor intervallet [-2.21;19.21] ◆

2.3.3 Median og kvartilafstand.

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 2.8: Median

Find medianen af tallene 6, 17, 7, 13, 5, 3.

Løsning:

Håndregning: Ordnet i rækkefølge: 3, 5, 6, 7 13, 17.

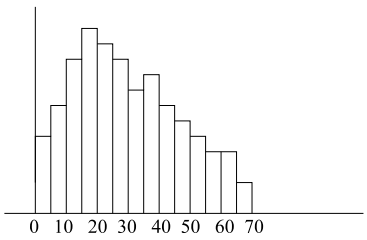
Excel: Som eksempel 2.6 nu vælges blot median

MEDIAN(A1:A6)	6,5

Medianen kaldes også for **50% fraktilen**, fordi den brøkdel (fraktil) der ligger under medianen er ca. 50% .

Er median og gennemsnit nogenlunde lige store fordeler tallene sig nogenlunde symmetrisk omkring middelværdien.

Er medianen mindre end gennemsnittet er der muligvis tale om en "højreskæv" fordeling som har den "lange" hale til højre. (se figuren)



At man eksempelvis i lønstatistikker 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.

Kvartilafstand.

Hvis fordelingen ikke er rimelig symmetrisk, er medianen det bedste skøn for en midterværdi, og kvartilafstanden kan være et mål for spredningen.
I den tidligere omtalte lønstatistik[1] findes bl.a. følgende tal, idet de to sidste kolonner er vor bearbejdning af tallene.

		Løn pr. præsteret time					
nr		gennemsnit $\bar{x}$	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 .

Eksempel 2.9 Kvartil

Find kvartiler og median af de 12 tal 7 , 9 , 11, 3 , 16, 12, 15, 8, 2, 18, 22, 10

Løsning:

Skal man kun have kvartiler.

Data indtastes i eksempelvis søjle A1 til A12 ▶ Tryk på f_x =▶statistik▶ På rullemenu vælges "Kvartil.medtag"

Der fremkommer en tabel med anvisning på, hvordan den skal udfyldes.

Resultat : 1. kvartil 7.75 3 kvartil 15.25

E	F	G
7		
9		
11	KVARTIL.MEDTAG(E1:E12;1)	7,75
3		
16	KVARTIL.MEDTAG(E1:E12;3)	15,25
12		
15		
8		
2		
18		
22		
10		

Ønsker man mange oplysninger herunder medianer
Data ▶ Dataanalyse ▶ Beskrivende statistik ▶ udfyld inputområde ▶ Resumestatistik

Det ses bl.a. at medianen er 10.5

[1] jævnfør statistisk årbog 2005 tabel 144 eller se www.statistikbanken.dk under løn\lønstatistik for den offentlige sektor \løn 32

OPGAVER

Opgave 2.1.
Følgende tabel angiver for et udvalgt antal lande oplysning om middellevetid for befolkningen og indbyggerantal.

Land	Middellevetid	Indbyggertal i millioner
Australien	80.3	19.9
Canada	80.0	32.5
Danmark	77,5	5.5
Frankrig	79.4	60.4
Marokko	70.4	32.2
Polen	74.2	38.6
Sri Lanka	72.9	19.9
USA	77.4	293.0

1) Indskriv ovenstående tabel i Excel, hvor landene er opskrevet alfabetisk.

Benyt Excel til

1) at ordne landene efter middellevetid (længst levetid først), og afbild dem grafisk.
2) tegn i et koordinatsystem to kurver, som angiver såvel landenes størrelse som middellevetid

Opgave 2.2
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 (som kan findes som excel-fil på adressen www.larsen-net.dk) 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) Find det største og mindste tal blandt observationerne.
2) Tegn et histogram, hvor intervallerne er lige brede, og hvor et af endepunkterne er tallet 80.
3) Beregn gennemsnit, spredning og median.
4) Vurder på baggrund heraf om fordelingen er nogenlunde symmetrisk (normalfordelt).
5) Angiv hvor stor en procentdel af bilerne, der kører over 80 km/h.

Opgave 2.3

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 (som kan findes som excel-fil på adressen www.larsen-net.dk) 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 karakteristiske værdier.

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.

Opgave 2.4

Den følgende tabel (som kan findes som excel-fil på adressen www.larsen-net.dk) viser vægtene (i kg) af 80 kaniner.

2.90	2.55	2.95	2,70	3.20	2.75	3.20	2.85	2.60	2.90	2.85	2.70	2.80	2.55	3.10	2.90
2.60	2.45	2.65	3.15	3.40	2.90	3.00	2.50	2.95	3.00	3.25	2.80	2.70	2.60	2.80	2.70
2.45	2.70	2.65	2.95	2.80	2.85	2.70	2.95	3.05	2.65	2.70	2.70	3.00	2.80	2.70	3.00
2.75	2.75	2.85	2.70	2.95	2.75	2.70	2.65	3.05	2.90	3.00	2.75	2.60	3.00	3.15	2.60
2.60	2.80	2.45	2.95	2.65	2.90	2.95	2.90	2.95	2.75	2.75	2.80	3.00	2.50	3.00	3.15

1) Foretag en vurdering af, om fordelingen er nogenlunde symmetrisk (normalfordelt) ved
   a) at tegne et histogram
   b) at beregne karakteristiske værdier
2) Angiv hvor stor en procent af kaninerne, der "approksimativt" overstiger en vægt på 3 kg.

3 STOKASTISK VARIABEL

3.1 SANDSYNLIGHED

Statistik bygger på sandsynlighedsteorien, som giver metoder til at finde, hvor stor chancen (sandsynligheden) er for at et bestemt resultat af et eksperiment forekommer.

DEFINITION af tilfældigt eksperiment. *Et eksperiment som kan resultere i forskellige udfald, selv om eksperimentet gentages på samme måde hver gang, kaldes et tilfældigt eksperiment (engelsk : random experiment)*

Det er karakteristisk for tilfældige eksperimenter, at man kan afgrænse en mængde kaldet eksperimentets **udfaldsrum** *U*, der indeholder de mulige **udfald**. Derimod kan man ikke forudsige, hvilket udfald der vil indtræffe ved udførelsen af eksperimentet.
Består eksperimentet eksempelvis i kast med en terning er udfaldsrummet $U = \{1, 2, 3, 4, 5, 6\}$, men man kan ikke forudsige udfaldet af næste kast (eksperiment). Selv om man 4 gange i træk har fået udfaldet "øjental 1", kan man ikke forudsige, hvilket udfald der indtræffer næste gang. Resultatet af 5. kast afhænger ikke af resultaterne af de foregående 4 spil. Man siger, at eksperimenterne er "**statistisk uafhængige**" (en præcis definition ses i kapitel 11).

Som eksempler på tilfældige eksperimenter kan nævnes:

a) Ét kast med en mønt. Udfaldsrum $U = \{\text{Plat}, \text{Krone}\}$.

b) Fremstilling af et parti levnedsmiddel og måling af det procentvise indhold af protein.
U = mængden af reelle tal fra 0 til 100.

c) Udtage en stikprøve på 400 elektroniske komponenter af en dagsproduktion og optælling af antallet af defekte komponenter. $U = \{0, 1, 2, 3, 4, 5, \ldots, 400\}$

d) Udtagning af et tilfældigt TV-apparat fra en dagsproduktion af TV-apparater og optælling af antallet af loddefejl. U = mængden af positive hele tal.

En hændelse er en delmængde af et eksperiments udfaldsrum.
Eksempelvis er *A*: "At få et lige øjental" en hændelse ved kast med en terning.
Hændelsen *A* siges at indtræffe, hvis et udfald fra *A* forekommer.

Sandsynlighedsbegrebet tager udgangspunkt i det i kapitel 1 omtalte begreb "relativ hyppighed".

DEFINITION af relativ hyppighed for hændelse A. *Gentages et eksperiment n gange, og forekommer hændelsen A netop* n_A *gange af de n gange, er A's relative hyppighed* $h(A) = \frac{n_A}{n}$

Lad eksempelvis eksperimentet være kast med en terning og hændelsen *A* være at få et lige øjental. Kastes terningen 100 gange og bliver resultatet et lige øjental 45 af de 100 gange er $h(A) = 0.45$.
Det er en erfaring, at øges antallet af gentagelser af eksperimentet, vil den relative hyppighed af hændelsen *A* stabilisere sig. Når *n* går mod ∞ ,vil den relative hyppighed erfaringsmæssigt nærme sig til en grænseværdi ("**de store tals lov**").

Ved sandsynligheden for A som benævnes $\boldsymbol{P(A)}$ forstås denne grænseværdi. (P = probability)

Da definitionen af sandsynlighed bygger på relativ hyppighed, er det naturligt, at det for ethvert par af hændelser A og B i udfaldsrummet U skal gælde :
$0 \le P(A) \le 1$, $P(U) = 1$ og P(enten A eller B) = P(A) + P(B) (skrives kort $P(A \cup B) = P(A) + P(B)$) forudsat A og B ingen elementer har fælles (er disjunkte).
(en mere generel regel findes i kapitel 8)
De 3 regler kaldes sandsynlighedsregningens aksiomer.
I kapitel 8 udledes på dette grundlag en række regler for regning med sandsynligheder.

Eksempel 3.1 Regler
Lad A = at få et ulige øjental ved et kast med en terning
B = at få en sekser ved et kast med en terning
Find sandsynligheden for enten at få et ulige øjental eller en sekser (evt. begge dele) ved kast med en terning.
Løsning:
$P(A) = \frac{1}{2}$. $P(B) = \frac{1}{6}$ $\qquad P(A \cup B) = P(A) + P(B) = \frac{1}{2} + \frac{1}{6} = \frac{2}{3}$ ◆

3.2 STOKASTISK VARIABEL

Ethvert statistisk problem må det på en eller anden måde være muligt at behandle talmæssigt. Betragtes 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.

DEFINITION af stokastisk variabel (engelsk: random variable). *En stokastisk variabel (også kaldet statistisk variabel) er en funktion, som tilordner et reelt tal til hvert udfald i udfaldsrummet for et tilfældigt eksperiment.*

En stokastisk variabel betegnes med et stort bogstav såsom X, mens det tilsvarende lille bogstav x betegner en mulig værdi af X.
Ved en **diskret** variabel (eller tællevariabel) forstås en variabel, hvis mulige værdier udgør en endelig eller tællelig mængde.
Er eksempelvis eksperimentet "udtagning af en kasse med 100 møtrikker, ud af en løbende produktion af kasser", kunne den stokastiske variabel X være defineret som " antal defekte møtrikker i kassen".
X er en diskret variabel, da den kun kan antage heltallige værdier fra 0 til 100.

Vi vil i senere afsnit behandle diskrete variable.

Ved en **kontinuert stokastisk variabel** forstås en stokastisk variabel, hvis mulige værdier er alle reelle tal i et vist interval.
Et eksempel kunne være eksperimentet "anvendelse af en ny metode til fremstilling af et produkt". Her kunne den stokastiske variabel Y være det målte procentvise udbytte ved forsøget.
Y er en kontinuert variabel, da den kan antage alle værdier fra 0% til 100%.

3.3 TÆTHEDSFUNKTION FOR KONTINUERT STATISTISK VARIABEL

Eksempel 3.2. Kontinuert stokastisk variabel
I menneskers led udskiller den inderste hinde en "ledvæske" som "smører" leddet. For visse ledsygdomme kan koncentrationen af brintioner (pH) i denne væske tænkes at have betydning. Som led i en nordisk medicinsk undersøgelse af en bestemt ledsygdom udtog man blandt samtlige patienter der led af denne sygdom tilfældigt 75 patienter og målte pH i ledvæsken i knæet. Resultaterne findes i eksempel 2.5

Population og stikprøve. Samtlige indbyggere i Norden med denne sygdom udgør **populationen**. Da det er ganske uoverkommeligt at undersøge alle, udtages en **stikprøve** på 75 patienter.

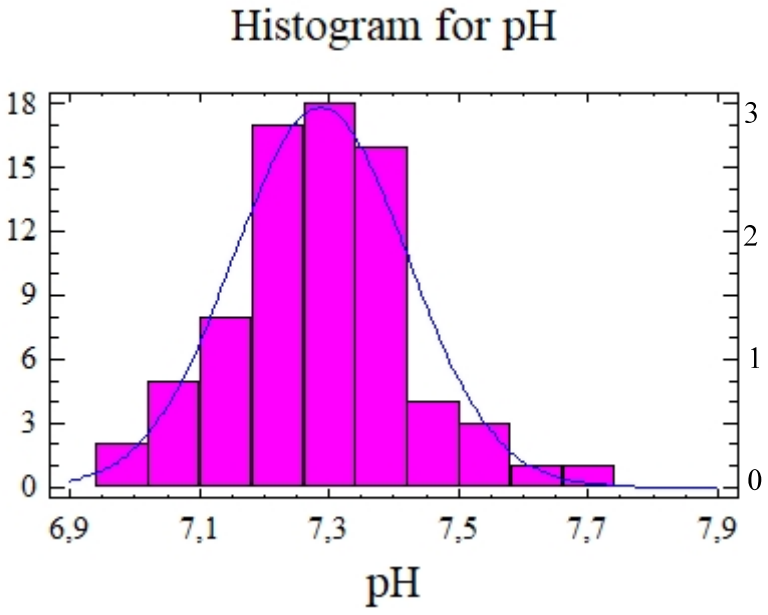
Det er målet ved hjælp af statistiske metoder på basis af en stikprøve at sige noget generelt om populationen.

Histogram. For at få et overblik over et større datamateriale, vil man sædvanligvis starte med at tegne et histogram. Hvorledes dette gøres fremgår af eksempel 2.5.
I skemaet ses resultatet af en opdeling i 10 klasser med en bredde på 0.08.
Endvidere er der beregnet en søjle ved at dividere den relative hyppighed med intervallængden.

Klasser	Antal n	Relativ hyppighed $\frac{n}{75}$	Skalering $\frac{n}{75 \cdot 0.08}$
]6.94 - 7.02]	2	0.0267	0.3333
]7.02 - 7.10]	5	0.0667	0.8333
]7.10 - 7.18]	8	0.1067	1.3333
]7.18 - 7.26]	17	0.2267	2.8333
]7.26 - 7.34]	18	0.2400	3.0000
]7.34 - 7.42]	16	0.2133	2.6667
]7.42 - 7.50]	4	0.0533	0.6667
]7.50 - 7.58]	3	0.0400	0.5000
]7.58 - 7.66]	1	0.0133	0.1667
]7.66 - 7.74]	1	0.0133	0.1667

Vi får det nedenfor tegnede histogram

Dette viser et "klokkeformet histogram", hvor der er flest tal fra 7.19 til 7.42, og derefter falder antallet 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. Hvis man tænker sig antallet af forsøg stiger (for eksempel undersøger hele populationen på måske 1 million nordiske knæ), 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 (indtegnet på grafen).

Hvis man benytter den salderede skala fra skemaet, som også er afsat på højre side af tegningen, vil arealet af hver søjle være den relative hyppighed, og for den 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 (p værdien) har en **tæthedsfunktion f(x)** hvis graf er den ovenfor nævnte kontinuerte kurve.
Da arealet under en kontinuert kurve beregnes ved et bestemt integral, følger heraf følgende definition:

DEFINITION af tæthedsfunktion f(x) for kontinuert variabel X.

$P(a \le X \le b) = \int_a^b f(x)dx$ *for ethvert interval af reelle tal* $[a\,;b]$

$\int_{-\infty}^{\infty} f(x)dx = 1, \qquad f(x) \ge 0$ for alle x

Bemærk, at for kontinuerte variable er
$P(a \le X \le b) = P(a < X \le b) = P(a \le X < b) = P(a < X < b).$

Et eksempel på en tæthedsfunktion for en kontinuert variabel er den i næste kapitel beskrevne normalfordeling.

Måleresultater vil sædvanligvis være værdier af normalfordelte variable, så en rimelig hypotese for den i eksempel 3.2 angivne kontinuerte stokastiske variabel X = p er således, at den er normalfordelt. Dette bestyrkes af at grafen for sådanne netop er klokkeformede .

Det er væsentlig at finde en central værdi i populationen, samt angive et spredningsmål
Disse angives i de følgende kapitler for de konkrete funktioner, der behandles.

Generelt gælder følgende definitioner
DEFINITION af middelværdi for kontinuert variabel. *Middelværdi for en kontinuert variabel X med tæthedsfunktion f (x) benævnes* μ *eller E (X) og er defineret som* $\mu = E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx$

◆

DEFINITION af varians og spredning for kontinuert variabel. *Variansen for en kontinuert variabel X med tæthedsfunktion f (x) benævnes* σ^2 *eller V(X) og er defineret som* $\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)dx$

Spredningen (engelsk: standard deviation) for en diskret variabel X med tæthedsfunktion f(x) benævnes σ *og er defineret som* $\sigma = \sqrt{V(X)}$ ◆

Eksempel 3.3 Kontinuert stokastisk variabel.

Lad der være givet følgende funktion: $f(x)=\begin{cases}\frac{3}{8}\cdot x^2 & \text{for } 0\le x<2\\ 0 & \text{ellers}\end{cases}$.

a) Vis, at $\int_{-\infty}^{\infty} f(x)dx=1$

I det følgende antages, at $f(x)$ er tæthedsfunktion for en kontinuert stokastisk variabel X.

b) Skitser grafen for f.

c) Beregn middelværdi og spredning for X.

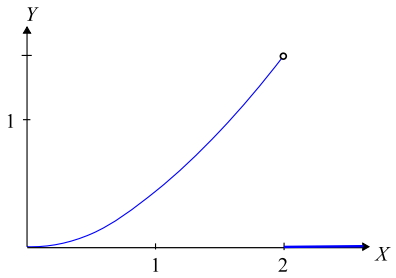


Fig.3.1 *Tæthedsfunktion*

Løsning:

a) $\int_{-\infty}^{\infty} f(x)dx=\int_0^2 \frac{3}{8}x^2dx=\left[\frac{x^3}{8}\right]_0^2=1$.

b) Grafen, som er en del af en parabel, ses på Fig 3.1.

c) $\mu=E(X)=\int_{-\infty}^{\infty} x\cdot f(x)dx=\int_0^2 x\cdot\frac{3}{8}x^2dx=\left[3\frac{x^4}{32}\right]_0^2=\underline{\underline{\frac{3}{2}}}$.

$$V(X)=\int_{-\infty}^{\infty} x^2\cdot f(x)dx-\mu^2=\int_0^2 x^2\cdot\frac{3}{8}x^2dx-\left(\frac{3}{2}\right)^2=\left[3\frac{x^5}{40}\right]_0^2-2.25=\underline{\underline{0.15}}.\qquad \sigma(X)=\sqrt{0.15}=\underline{\underline{0.387}}.$$

◆

Fordelingsfunktion. I visse situationer er det en fordel at betragte den kontinuerte variabels fordelingsfunktion $F(x)$

DEFINITION af fordelingsfunktion F(x) for kontinuert variabel.

Fordelingsfunktionen for en kontinuert variabel X er defineret ved $F(x)=P(X\le x)=\int_{-\infty}^{x} f(x)dx$

◆

DEFINITION af p-fraktil x_p . *Lad p være et vilkårligt tal mellem 0 og 1.*
Ved p-fraktilen eller 100 p % fraktilen forstås det tal x_p *, for hvilket det gælder, at*

$F(x_p)=P(X\le x_p)=p \quad (=\int_0^{x_p} f(x)dx)$

◆

Særlig ofte benyttede fraktiler er 50% fraktilen, som kaldes **medianen** (eller 2. kvartil), 25 % fraktilen, som kaldes **nedre kvartil** (eller 1. kvartil) og 75% fraktilen, som kaldes **øvre kvartil** (eller 3. kvartil).

Eksempel 3.4. Fordelingsfunktion for kontinuert variabel.
For den i eksempel 3.3 angivne kontinuerte variabel X med tæthedsfunktion $f(x)$ ønskes fundet:
1) Fordelingsfunktionen $F(x)$
2) Medianen

Løsning:

1) $$F(x)=\int_{-\infty}^{x} f(x)dx=\begin{cases}\int_{-\infty}^{x} 0\,dx= & 0 \quad \text{for } x<0\\ 0+\int_0^x \frac{3}{8}x^2dx=\left[\frac{x^3}{8}\right]_0^x=\frac{x^3}{8} & \text{for } 0\le x\le 2\\ 0+\frac{2^3}{8}+\int_2^x 0dx= & 1 \quad \text{for } x>2\end{cases}.$$

2) Medianen er bestemt ved $F(x)=0.5\Leftrightarrow\frac{x^3}{8}=0.5\Leftrightarrow x^3=4\Leftrightarrow\underline{\underline{x=1.59}}$.

◆

3.4 LINEARKOMBINATION AF STOKASTISKE VARIABLE

Vi betragter i dette afsnit flere stokastiske variable.
Eksempel 3.5 vil blive benyttet som gennemgående eksempel

Eksempel 3.5. To variable.
Insektpulver sælges i papkartoner. Lad den stokastiske variable X_1 være vægten af pulveret, mens X_2 er vægten af papkartonen. I middel fyldes der 500 gram insektpulver i hver karton med en spredning på 5 gram. Kartonen vejer i middel 10 gram med en spredning på 1.0 gram.
$Y = X_1 + X_2$ er da bruttovægten.
1) Find middelværdien af Y
2) Find spredning af Y.

◆

Mere generelt haves:
Lad $X_1, X_2, \ldots, X_n$ være n stokastiske variable.
Ved en **linearkombination** af disse forstås
$Y = a_0 + a_1 \cdot X_1 + a_2 \cdot X_2 + \ldots + a_n \cdot X_n$, hvor $a_0, a_1, a_2, \ldots, a_n$ er konstanter.

Der gælder følgende
Linearitetsregel: $E(Y) = a_0 + a_1 \cdot E(X_1) + a_2 \cdot E(X_2) + \ldots + a_n \cdot E(X_n)$.

I eksempel 3.5 synes det rimeligt at antage, at vægten af pulveret og vægten af papkartonen er **uafhængige** (påfyldningen kan tænkes at ske maskinelt, uden at den er afhængig på nogen måde af hvilken vægt, kartonen tilfældigvis har).

Man kan vise (se eventuelt på hjemmesiden "Larsen-net.dk" denne bog version 8 kapitel 9, for en mere udførlig behandling af uafhængighed, bevis for kvadratregel m.m.), at hvis $X_1, X_2, \ldots, X_n$ er **statistisk uafhængige,** gælder

Kvadratregel for statistisk uafhængige variable:
$V(Y) = a_1^2 \cdot V(X_1) + a_2^2 \cdot V(X_2) + \ldots + a_n^2 \cdot V(X_n)$.

Eksempel 3.5. (fortsat) **To variable.**
Spørgsmål 1: $E(Y) = E(X_1) + E(X_2) = 500 + 10 = \underline{\underline{510 \text{ gram.}}}$
Spørgsmål 2: $V(Y) = V(X_1) + V(X_2) = 5^2 + 1^2 = 26.$ $\sigma(Y) = \sqrt{26} = \underline{\underline{5.1}}$ gram. ◆

Ensfordelte uafhængige variable.
Lad os antage, at vi uafhængigt af hinanden og under de samme betingelser udtager n elementer fra en population med middelværdi μ og spredning σ. Lad X_1 være den stokastiske variabel, der er resultatet af første udtagning af et element i stikprøven, X_2 være den stokastiske variabel, der er resultatet af anden udtagning, osv.

$X_1, X_2, \ldots, X_n$ vil da være **ensfordelte** uafhængige stokastiske variable, dvs. have samme fordeling med middelværdi μ og spredning σ.

Eksempel 3.6. Ensfordelte variable

Bruttovægten af det i eksempel 3.5 nævnte karton insektpulver havde middelvægten 510 g med en spredning på 5.1 g.
Vi udtager nu tilfældigt og uafhængigt af hinanden 10 pakker insektpulver.

a) Hvad bliver i middel den samlede vægt af de 10 kartonner
b) Hvad bliver i middel spredningen på den samlede vægt af de 10 kartoner

Løsning:
Lad X_1 være vægten af karton 1, X_2 være vægten af karton 2 osv. X_{10} være vægten af karton 10.
$Y = X_1 + X_2 + \ldots + X_{10}$ er da vægten af alle 10 kartonner.

a) $E(Y) = E(X_1)+E(X_2)+ \ldots +E(X_{10}) = 10 \cdot 510 = \underline{\underline{5100 \text{ g}}}$
b) $V(Y) = V(X_1)+V(X_2)+ \ldots +V(X_{10}) = 10 \cdot (5.1)^2 = 260.1$ g

$$\sigma(Y) = \sqrt{260.1} = \underline{\underline{16.13}}$$

Bemærk: En almindelig fejl er her, at man tror, at $Y = 10 \cdot X$ og dermed $V(Y) = 10^2 \cdot V(X) = 2600$
Vi har her at gøre med 10 **ensfordelte uafhængige** variable, og ikke 10 · vægten af 1 karton.

◆

For ensfordelte uafhængige stokastiske variable gælder:

SÆTNING 3.1 (*middelværdi og spredning for stikprøves gennemsnit*)

$$E(\overline{X}) = \mu \text{ og } \sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}, \text{ hvor } \overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Bevis: Af linearitetsreglen fås $E(\overline{X}) = E\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) = \frac{1}{n}\left(E(X_1) + E(X_2) + \ldots + E(X_n)\right) = \mu$

Af kvadratreglen fås $V(\overline{X}) = V\left(\frac{X_1 + X_2 + \ldots X_n}{n}\right) = \frac{1}{n^2}\left(V(X_1) + V(X_2) + \ldots + V(X_n)\right) = \frac{n \cdot \sigma^2}{n^2} = \frac{\sigma^2}{n}$.

◆

Eksempel 3.7. Spredning på gennemsnit (eksempel 3.5 fortsat)

Hvis der udtages 5 kartoner insektpulver, hvad vil da være spredningen på gennemsnittet af vægten af insektpulveret .

Løsning:
Da spredningen på 1 karton er 5.1 gram, vil spredningen på gennemsnittet af 5 kartoner være

$$\sigma(X) = \frac{\sigma}{\sqrt{n}} = \frac{5.1}{\sqrt{5}} = \underline{\underline{2.28}}$$

◆

Opgave 3.1

Vægten af en (tilfældigt udvalgt) tablet af en vis type imod hovedpine har middelværdien $\mu = 0.65$ g og spredningen $\sigma = 0.04$ g

a) Beregn middelværdi og spredning af den sammenlagte vægt af 100 (tilfældigt udvalgte) tabletter
b) På basis af de 100 tabletter ønskes spredningen på gennemsnittet beregnet.

4 NORMALFORDELINGEN

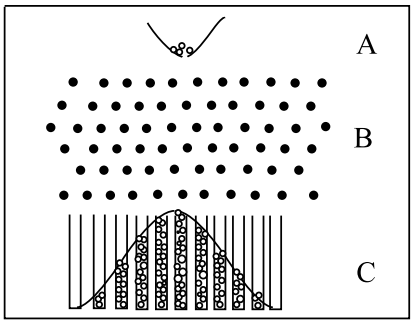
4.1 INDLEDNING

Lad os som eksempel tænke os et kemisk forsøg, hvor vi måler udbyttet af et stof A. Selv om vi gentager forsøget ved anvendelse af den samme metode og i øvrigt søger at gøre forsøgsbetingelserne så ensartet som muligt, varierer udbyttet dog fra forsøg til forsøg. Disse variationer fra den ene forsøg til det næste må skyldes forhold vi ikke kan styre. Det kan skyldes små ændringer i temperaturen, i luftens relative fugtighed, vibrationer under fremstillingen, små forskelle i de anvendte råmaterialer (kornstørrelse, renhed), forskelle i menneskelig reaktionsevne osv. Hvis ingen af disse variationsårsager er dominerende, der er et stort antal af dem, de er uafhængige og lige så godt kan have en positiv som en negativ indvirkning på resultatet, så vil den totale fejl sædvanligvis approksimativt være fordelt efter den såkaldte **normalfordeling.** (også kaldet Gauss-fordelingen)

Som illustration af dette kan anvendes Galtons apparat.

Eksempel 4.1. Eksperiment med et Galton-apparat.

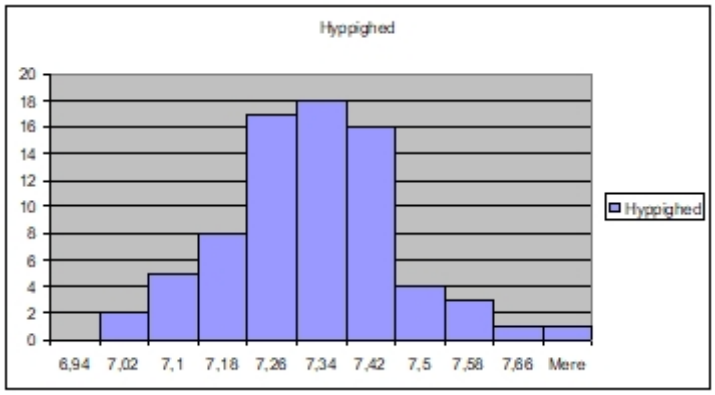
På den anførte figur er skitseret et Galton-apparat.
A er en tragt; B er sømrækker, hvor sømmene i en underliggende række er anbragt midt ud for mellemrummene mellem sømmene i den overliggende række; C er opsamlingskanaler.



Lader man mange kugler passere gennem tragten A ned gennem sømrækkerne B til opsamlingskanalerne C, vil man konstatere, at de enkelte kugler nok bliver tilfældigt fordelt i opsamlingskanalerne, men at kuglernes samlede fordeling giver et mønster, som gentages, hver gang man udfører eksperimentet. Fordelingen er hver gang med tilnærmelse en klokkeformet symmetrisk fordeling som skitseret på tegningen, noget som er karakteristisk for normalfordelingen.

Galton-apparatet illustrerer, hvorfor man så ofte antager, at måleresultater er værdier af en normalfordelt variabel: Hver sømrække repræsenterer en faktor, hvis niveau det ikke er muligt at holde konstant fra måling til måling, og sømrækkernes påvirkning af kuglens bane symboliserer den samlede virkning, som de ukontrollerede faktorer har på størrelsen af den målte egenskab. ◆

En anden illustration af under hvilke omstændigheder en normalfordelt variabel kan forekomme i praksis så vi i kapitel 2 eksempel 2.5 hvor man på 75 mennesker med en bestemt ledsygdom målte pH i knæleddet.
Histogrammet som er gentaget nedenfor har et klokkeformet udseende, som kraftigt antyder, at den kontinuerte stokastiske variabel X = pH er normalfordelt.



I den teoretiske statistik giver den centrale grænseværdisætning en forklaring på, hvorfor normalfordelingen er en god model ved mange anvendelser.

Den centrale grænseværdi siger (løst sagt), at selvom man ikke kender fordelingen for de n ensfordelte stikprøvevariable $X_1, X_2, \ldots, X_n$, så vil gennemsnittet $\overline{X}$ være approksimativt normalfordelt blot n er tilstrækkelig stor (i praksis over 30) .

4.2 DEFINITION OG SÆTNINGER OM NORMALFORDELING

Definition af normalfordeling $n(\mu, \sigma)$

Nornalfordelingen er sandsynlighedsfordelingen for en kontinuert stokastisk variabel X med tæthedsfunktionen f(x) bestemt ved $f(x) = \frac{1}{\sqrt{2 \cdot \pi} \cdot \sigma} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ for ethvert x

Den har middelværdi μ og spredning σ
Grafen er klokkeformet og symmetrisk om linien x = μ.

◆

At $f(x)$ virkelig er en tæthedsfunktion med de angivne egenskaber vises i "Supplement til statistiske grundbegreber afsnit 2.A"

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

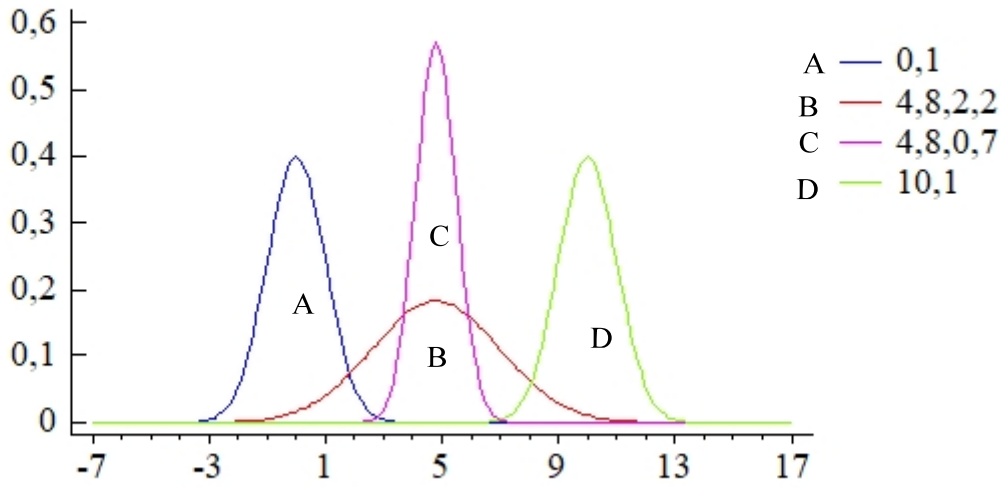


Fig 4.1 *Forskellige normalfordelinger*

Det ses, at tæthedsfunktionerne er klokkeformede, og at et interval på $[\mu - 3 \cdot \sigma; \mu + 3 \cdot \sigma]$ indeholder stort set hele sandsynlighedsmassen.

Vi nævner uden bevis følgende sætning:

SÆTNING 4.1 Additionssætning for linearkombination af normalfordelte variable.
Er Y en linearkombination af n stokastisk uafhængige , normalfordelte variable, vil Y også være normalfordelt.

Kendes middelværdi og spredning for de n normalfordelte variable , kan man ved anvendelse af linearitetsregel og kvadratregel finde Y's middelværdi og spredning.

Endvidere følger det af additionssætningen, og sætning 3.1, at gennemsnittet $\bar{x}$ er normalfordelt med en spredning på $\frac{\sigma}{\sqrt{n}}$.

Normeret normalfordeling

Af særlig interesse er den såkaldte normerede normalfordeling.
Den er bestemt ved at have middelværdien 0 og spredningen 1.
Grafen for den er tegnet som graf A i figur 4.1
Den kaldes sædvanligvis *U* eller Z og dens fordeling **U- eller Z-fordelingen** . Dens tæthedsfunktion benævnes φ og dens fordelingsfunktion Φ.
Specielt vil dens p - fraktil z_p indgå i adskillige formler i de næste afsnit.

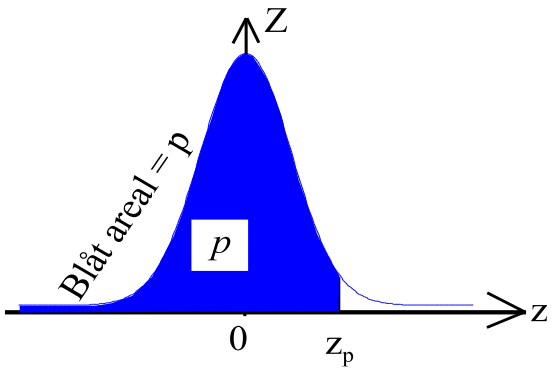


Fig 4.2 *Normeret normalfordeling*

En vigtig sammenhæng mellem fraktiler for *X* og fraktiler for *Z* er følgende

(4.1) $x_p = z_p \cdot \sigma + \mu$

0 Beviset for denne relation indgår i beviset for den følgende sætning, som også viser, at man kan overføre en vilkårlig normalfordeling til den normerede normalfordeling.
Det er derfor nok at lave en tabel over den normerede normalfordeling.
Dette er det man udnytter, hvis man ikke har rådighed over et program, der som beskrevet i afsnit 4.3 direkte kan beregne værdierne.
Der gælder følgende

SÆTNING 4.2. (normering af normalfordeling). *Når X er normalfordelt* $n(\mu, \sigma)$
er den variable $Z = \frac{X-\mu}{\sigma}$ *normalfordelt* $n(0,1)$, *og der gælder*

$$P(X \le b) = \Phi\left(\frac{b-\mu}{\sigma}\right) \quad \text{og} \quad P(a<X\le b)=\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right).$$

Endvidere gælder $x_p = z_p \cdot \sigma + \mu$

Bemærk, at det for de to formler er ligegyldigt, om ulighederne er med eller uden lighedstegn.

Bevis:
At Z også er normalfordelt vises ikke her.

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

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

Z har derfor middelværdi 0 og spredning 1.

Endvidere fås $P(X\le b)=P\left(\frac{X-\mu}{\sigma}\le\frac{b-\mu}{\sigma}\right)=P\left(Z\le\frac{b-\mu}{\sigma}\right)=\Phi\left(\frac{b-\mu}{\sigma}\right)$ og

$$P(a<X\le b)=P\left(\frac{a-\mu}{\sigma}<Z\le\frac{b-\mu}{\sigma}\right)=\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)$$

Bevis for $x_p=z_p\cdot\sigma+\mu$: $P(X\le x_p)=p\Leftrightarrow\Phi\left(\frac{x_p-\mu}{\sigma}\right)=p\Leftrightarrow\frac{x_p-\mu}{\sigma}=z_p\Leftrightarrow x_p=z_p\cdot\sigma+\mu$

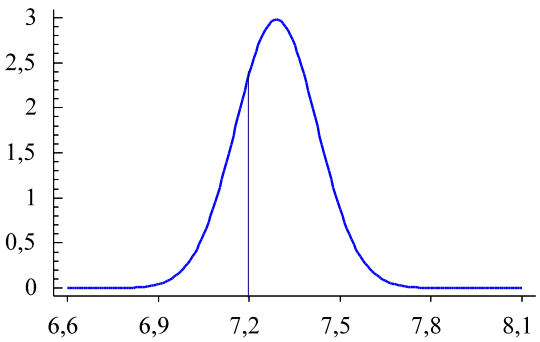
4.3. BEREGNING AF SANDSYNLIGHEDER

Stikprøves gennemsnit og spredning.

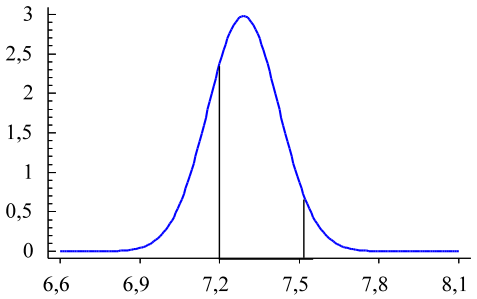
Ofte er middelværdien μ og spredningen σ ukendt i en foreliggende normalfordeling. I så fald erstattes fordelingen $n(\mu,\sigma)$ i praksis med en approksimerende fordeling $n(\bar{x},s)$, såfremt der foreligger et rimelig stort antal observationer fra den givne fordeling.
På basis af den i eksempel 1.5 angivne stikprøve på 75 patienter beregnes et gennemsnit af pH værdierne på $\bar{x}=\frac{546.52}{75}=7.2868$ og en s værdi på $s=\sqrt{\frac{SAK}{n-1}}=0.134355$.

Vi vil altså antage, at pH værdierne er approksimativt normalfordelt n (7.29, 0.134).

Ønsker vi at benytte ovenstående normalfordeling n (7.29, 0.134) til at finde sandsynligheden for, at pH er mindre end 7.2, er denne sandsynlighed lig med arealet af det skraverede areal under tæthedsfunktionen.



Ønsker vi tilsvarende at beregne sandsynligheden for, at pH ligger mellem 7.2 og 7.5 er sandsynligheden lig med det skraverede areal under kurven på omstående figur.



Eksempel 4.2. Beregning af normalfordelte sandsynligheder

Lad X være normalfordelt $n(\mu, \sigma)$, hvor $\mu = 7.29$ og $\sigma = 0.134$.

1) Find $P(X \leq 7.2)$
2) Find $P(7.2 \leq X \leq 7.5)$
3) Find $P(X > 7.6)$
4) Find 90% fraktilen $x_{0.9}$

Løsning:

Man finder de benyttede sandsynlighedsfordelinger ved at vælge

På værktøjslinien foroven: Tryk f_x ▶Vælg kategorien "Statistisk"▶Udfyld menu

A	B	C
1) P(X<=7.2)=	NORMAL.FORDELING(7,2;7,29;0,134;1)	0,250906
2) P(7.2<X<7.5)=	NORMAL.FORDELING(7,5;7,29;0,134;1)-NORMAL.FORDELING(7,2;7,29;0,134;1)	0,690556
3) P(X>7.6)=	1-NORMAL.FORDELING(7,6;7,29;0,134;1)	0,010349
4) x =	NORM.INV(0,9;7,29;0,134)	7,461728

2) Beregningen sker ved at beregne arealet fra $-\infty$ til 7.5 og derfra trække arealet fra $-\infty$ til 7.2
3) Da arealet under kurven er 1, fås $P(X \geq 7.6) = 1 - P(X < 7.6)$
4) Har man omvendt givet en sandsynlighed p = 0.9 og ønsker at finde den tilsvarende værdi x_p for hvilken $P(X \leq x_p) = 0.9$ betyder det man kender arealet 0.9 og skal finde x-værdien. Det svarer til at finde den (inverse) omvendte funktion af normalfordelingen.

◆

Eksempel 4.3. Kvalitetskontrol.

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 tolerancegrænserne 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.

a) Hvis X har en middelværdi på 89.6, hvad er så sandsynligheden for, at en kasse har en længde, der ligger indenfor tolerancegrænserne.
b) 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.

c) Hvad skal spredningen σ formindskes til, for at dette er opfyldt?

Hvis det er umuligt at ændre σ, kan man prøve at få ændret tolerancegrænserne.

d) Find de nye tolerancegrænser (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.

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

Løsning
Man finder de benyttede sandsynlighedsfordelinger på samme måde som i eksempel 4.2
Tryk f_x ▶Vælg kategorien "Statistisk"

a) $P(89.2 \le X \le 90.8) = P(X \le 90.8) - P(X \le 89.2) =$
NORMFORDELING(90,8;89,6;0,5;1) - NORMFORDELING(89,2;89,6;0,5;1)=0,7799

b) Middelværdien justeres til midtpunktet 90.0
$P(X > 90.8) + P(X < 89.2) = 1 - P(X \le 90.8) + P(X < 89.2) =$
1 -NORMFORDELING(90,8;90;0,5;1) - NORMFORDELING(89,2;90;0,5;1) = 0.1096

c) Da der ligger 5% udenfor intervallet, så må af symmetrigrunde 2,5% ligge på hver sin side af intervallet. Vi har følgelig, at vi skal finde spredningen σ så $P(X \le 89.2) = 0.025$

$$89.2 = z_{0.025} \cdot \sigma + 90 \Leftrightarrow \sigma = \frac{-0.8}{z_{0.025}} \Leftrightarrow \sigma = (-0,8) / NORN.INV(0,025;0;1) = 0.4082 .$$

c) Da der ligger 5% udenfor intervallet, så må af symmetrigrunde 2,5% ligge på hver sin side af intervallet. Vi har følgelig, at vi skal finde spredningen σ så $P(X \le 89.2) = 0.025$
Metode 1: Af relationen (4.1) fås

$$89.2 = z_{0.025} \cdot \sigma + 90 \Leftrightarrow \sigma = \frac{-0.8}{z_{0.025}} \Leftrightarrow \sigma = \frac{-0.8}{Normal\ Quantile(0.025)} = 0.4082 .$$

Metode 2: I celle A1 skrives en startværdi for σ eksempelvis 0,5.
▶I celle B1 skrives =▶f_x▶ NORMFORDELING▶udfyld menu med 89,2;90;A1;1▶
Data▶ What if analyse▶"Målsøgning
I "Angiv celle" skrives B1. I "Til Værdi" skrives 0,025. I "Ved ændring af celle" skrives A1.

A	B	C
0,408444	0,025077	NORMAL.FORDELING(89,2;90;A1;1)

Facit :0,408444

d) $P(90.0 - d < X < 90.0 + d) = 0.95 \Leftrightarrow P(X \le 90.0 - d) = 0.025$ og $P(X \le 90.0 + d) = 0.975$.
Vi får nedre grænse =NORMINV(0,025;90;0,5) = 89,02002 = 89.0
Øvre grænse =NORMINV(0,975;90;0,5) = 90,97998 = 91.0
e) Efter indtastning af de 12 tal i Excel i cellerne A1 til A12 trykkes på f_x
$\bar{X} =$ MIDDELV(A1:A12) = 89.97 og s = STDAFV(A1:A12) = 0.435

Eksempel 4.4. Additionssætning.

En boreproces fremstiller huller med en diameter X_1, der er normalfordelt med en middelværdi μ_1 og en spredning på 0.04. En anden proces fremstiller aksler med en diameter X_2, der er normalfordelt med en middelværdi μ_2 og en spredning på 0.03.
Antag, at $\mu_1 = 10.00$, og at $\mu_2 = 9.94$.
Find sandsynligheden for, at en tilfældig valgt aksel har en mindre diameter end en tilfældig valgt borehul.
Løsning:
$P(X_2 < X_1) = P(X_2 - X_1 < 0).$
Sættes $Y = X_2 - X_1$ er Y normalfordelt. $E(Y) = E(X_2) - E(X_1) = 9.94 - 10.00 = -0.06$
$V(Y) = 1^2 V(X_2) + (-1)^2 V(X_1) = 0.04^2 + 0.03^2 = 0.025 \quad \sigma(Y) = \sqrt{0.0025} = 0.05$
Excel: $P(X_2 < X_1) = P(Y < 0)$ = normalfordeling(0;-0,06;0,05;1)= 0.8849 = 88.49% ◆

OPGAVER

Opgave 4.1

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 $\mu = 2.1$ og $\sigma = 2.4$.
   Find $P(22.3 < X \le 27.8)$.

Opgave 4.2

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

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 4.3

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 4.4

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

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

Opgave 4.5

I et laboratorium lægges et nyt gulv.
Det forudsættes, at vægten Y der hviler på gulvet, er summen af vægten X_1 af maskiner og apparater og vægten X_2 af varer og personale, dvs. $Y = X_1 + X_2$
Da både X_1 og X_2 er sum af mange relativt små vægte, antages det, at de er normalfordelte.
Det antages endvidere at X_1 og X_2 er statistisk uafhængige.
Erfaringer fra tidligere gør det rimeligt at antage, at der gælder følgende middelværdier og spredninger (målt i tons):
$E(X_1) = 6.0$, $\sigma(X_1) = 1.2$, $E(X_2) = 3.5$, $\sigma(X_2) = 0.4$.

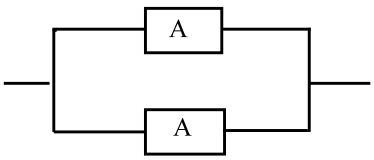
1) Beregn $E(Y)$ og $\sigma(Y)$.
2) Beregn det tal y_0 , som vægten Y med de ovennævnte forudsætninger kun har en sandsynlighed på 1% for at overskride.
3) Beregn sandsynligheden for, at vægten af varer og personale en tilfældig dag, efter at det nye gulv er lagt, er større end vægten af maskiner og apparater. (Vink: se på differensen X_2 - X_1)

Opgave 4.6

Ved fabrikation af et bestemt mærke opvaskemiddel fyldes vaskepulver i papkartoner.
I middel fyldes 4020 g pulver i hver karton, idet der herved er en spredning på 12 g.
Pulverfyldningen kan forudsættes ikke at afhænge af kartonernes vægt, der i middel er 250 g med en spredning på 5g.
Beregn sandsynligheden p for, at en tilfældig pakke opvaskemiddel har en bruttovægt mellem 4250 g og 4300 g.

Opgave 4.7

Et system er af sikkerhedsmæssige grunde opbygget af to apparater A, der er parallelforbundne (se figur) således, at systemet virker, så længe blot et af apparaterne virker.



Svigter et af apparaterne, startes reparation. Det antages, at reparationstiden er normalfordelt med middelværdien
μ_{rep} = 10 timer og spredning σ_{rep} = 3 timer.
I reparationstiden overbelastes den anden komponent, og det antages, at dens levetid fra reparationens start (approksimativt) er normalfordelt med middelværdi μ og spredning σ = 4 timer.

1) Find sandsynligheden for, at reparationen er afsluttet, inden den anden komponent fejler, hvis μ = 20 timer.
2) Hvor stor skal μ være, for at sandsynligheden for, at reparationen kan afsluttes før den anden komponent fejler, er mere end 99.9%?

Opgave 4.8

Vægten af en (tilfældig udvalgt) tablet af en vis type mod hovedpine har middelværdien 0.65 g og spredningen 0.04 g.

1) Beregn middelværdi og spredning af den sammenlagte vægt af 100 (tilfældigt udvalgte) tabletter.
2) Antag, at man benytter følgende metode til at fylde tabletter i et glas. Man placerer glasset på en vægt og fylder tabletter på, indtil vægten af tabletterne i glasset overstiger 65,3 g. Beregn sandsynligheden for, at glasset kommer til at indeholde mere end 100 tabletter (se bort fra vægtens fejlvisning).

5. Konfidensinterval for normalfordelt variabel

5.1 UDTAGNING AF STIKPRØVER

I langt de fleste i praksis forekomne tilfælde vil det bl.a. af tidsmæssige og omkostningsmæssige grunde være umuligt at foretage en totaltælling af hele populationen. Helt klart er dette ved afprøvningen ødelægger emnet (åbning af konservesdåser) eller populationen i princippet er uendelig (for at undersøge om en metode giver et større udbytte end et andet, udføres en række kemiske forsøg og her er der teoretisk ingen øvre grænse for antal delforsøg)
Som det senere vil fremgå kan selv en forholdsvis lille repræsentativ stikprøve give svar på væsentlige forhold omkring hele populationen.
Det er imidlertid klart, at en betingelse herfor er, at stikprøven er **repræsentativ**, dvs. at stikprøven med hensyn til den egenskab der ønskes er et "mini-billede" af populationen.
For at opnå det, foretager man en eller anden form for lodtrækning (kaldes **randomisering**).
Afhængig af problemet kan dette gøres på forskellig måde.

Simpel udvælgelse. Den enkleste form for stikprøveudtagning er, at man nummererer populationens elementer, og så **randomiserer** (ved lodtrækning, evt. ved at benyttet et program der generer tilfældige tal) udtager de N elementer der skal indgå i stikprøven.
Eksempel: For at undersøge om en ændring af vitaminindholdet i foderet for svin ændrede deres vægt, udvalgte man ved randomisering de svin, som fik det nye foder.

Stratificeret udvælgelse.
Under visse omstændigheder er det fordelagtigt (mindre stikprøvestørrelse for at opnå samme sikkerhed) at opdele populationen i mindre grupper (kaldet strada), og så foretage en simpel udvælgelse indenfor hver gruppe. Dette er dog kun en fordel, hvis elementerne indenfor hver gruppe er mere ensartet end mellem grupperne.
Eksempel: Ønsker man at spørge vælgerne om deres holdning til et politisk spørgsmål (f.eks. om deres holdning til et skattestop) kunne det måske være en fordel at dele dem op i indkomstgrupper (høj, mellem og lav) .

Systematisk udvælgelse.
Ved en såkaldt systematisk udvælgelse, vælger man at udtage hver k'te element fra populationen.
Eksempel: En detailhandler ønsker at måle tilfredsheden hos sine kunder. Der ønskes udtaget 40 kunder i løbet af en speciel dag.
Da man naturligvis ikke på forhånd kender de kunder der kommer i butikken, vælges en systematisk udvælgelse, ved at vælge hver 7'ende kunde der forlader butikken. Man starter dagen med ved lodtrækning at vælge et af tallene fra 1 til 7. Lad det være tallet 5. Man udtager nu kunde nr. $5,5+1\cdot 7=12,5+2\cdot 7=19,\ldots,5+39\cdot 7=278$. Derved har man fået valgt i alt 40 kunder.
Problemet er naturligvis, om tallet 7 er det rigtige tal. Hvis man får valgt tallet for stort, eksempelvis sætter det til 30, så vil en stikprøve på 40 kræve, at der er 1175 kunder den dag, og det behøver jo ikke at være tilfældet. Omvendt hvis tallet er for lille, så får man måske udtaget de 40 kunder i løbet af formiddagen, og så er stikprøven nok ikke repræsentativ, da man ikke får eftermiddagskunderne med.

Klyngeudvælgelse (Cluster sampling)
Denne metode kan med fordel benyttes, hvis populationen består af eller kan inddeles i delmængder (klynger) . Metoden består i, at man ved randomisering vælger et mindre antal klynger, som så totaltælles.

Eksempel: I et vareparti på 2000 emner fordelt på 200 kasser hver med 10 emner ønsker man en vurdering af fejlprocenten. I alt ønskes udtaget 50 emner.
Man udtager randomiseret 5 kasser, og undersøger alle emnerne i kasserne.

5.2. FORDELING OG SPREDNING AF GENNEMSNIT

Udtages en stikprøve fra en population er det jo for, at man ud fra stikprøven kan fortælle noget centralt om hele populationen.
I eksempel 1.5 var vi således interesseret i koncentrationen af brintioner (pH) i ledvæsken i knæet hos patienter, der led af denne sygdom.
Som led i en nordisk medicinsk undersøgelse udtog man blandt patienter der led af denne sygdom tilfældigt en stikprøve på 75.
På basis heraf beregnede man gennemsnittet af pH værdierne til $\bar{x}$ = 7.2868 og spredningen s = 0.134355 .
Man vil nu sige, at et estimat (skøn) for den "sande" middelværdi μ for hele populationen er 7.29 og den "sande" spredning" σ er 0.134.

Det er imidlertid klart, at disse tal er behæftet med en vis usikkerhed.
Havde vi valgt 75 andre patienter havde vi uden tvivl fået lidt andre tal.
Det er derfor ikke nok, at angive at den "sande" middelværdi er $\bar{x}$, vi må også angive et "usikkerhedsinterval".

For at kunne beregne et sådant interval er det nødvendigt at kende fordelingen.

Her spiller den tidligere nævnte centrale grænseværdisætning en vigtig rolle, idet den jo (løst sagt) siger, at selv om man ikke kender fordelingen af den kontinuerte stokastiske variabel, så vil ***gennemsnittet af værdierne i en stikprøve på n tal vil være tilnærmelsesvis normalfordelt, hvis blot n er tilstrækkelig stor (i praksis over 30).***

Dette er af stor praktisk betydning, idet det så ikke er så vigtigt, om selve populationen er normalfordelt. Ofte er det jo kun af interesseret at kunne forudsige noget om hvor middelværdien af fordelingen er placeret.

Endvidere fremgik det af sætning 3.1 , at spredningen på $\bar{x}$ er $\sigma(\bar{x}) = \frac{\sigma}{\sqrt{n}}$, hvor σ er spredningen på den enkelte værdi i stikprøven.
Heraf fremgår, at gennemsnittet kan man "stole" mere på end den enkelte måling, da den har en mindre spredning.

Eksempel 5.1. Fordeling af gennemsnit

Den tid, et kunde må venter i en lufthavn ved en check-in disk, er givet at være en stokastisk variabel med en ukendt fordeling. Man har dog erfaring for, at ventetiden i middel er på 8.2 minutter med en spredning på 3 minutter.

Udtages en stikprøve på 50 kunder, ønskes fundet sandsynligheden for, at den gennemsnitlige ventetid for disse kunder er mellem 7 og 9 minutter

Løsning:

Da antallet *n* i stikprøven på 50 er større end 30, kan vi antage at gennemsnittet er approksimativt normalfordelt med en middelværdi på 8.2 og en spredning på $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{50}} = 0.424$.

Vi har derfor

$P(7 < \bar{X} < 9) =$ NORMAL.FORDELING(9;8,2;0.424;1)-NORMAL.FORDELING(7;8,2;0.424;1)$= 0.9681 =$ 96.8%

5.3. KONFIDENSINTERVAL FOR MIDDELVÆRDI

5.3.1 Definition af konfidensinterval

Udtages en stikprøve fra en population er det jo for, at man ud fra stikprøven kan fortælle noget centralt om hele populationen.

Man vil eksempelvis beregne gennemsnittet $\bar{x}$ og angive det som et estimat (skøn) for den "sande" middelværdi μ for hele populationen

Det er imidlertid klart, at selv om et gennemsnit har en mindre spredning end den enkelte måling, så er det stadig behæftet med et vis usikkerhed

Det er derfor ikke nok, at angive at den "sande" middelværdi er $\bar{x}$, vi må også angive et "usikkerhedsinterval".

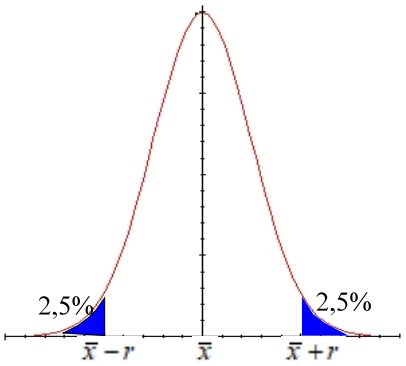
Et interval indenfor hvilket den "sande værdi" μ med eksempelvis 95% "sikkerhed" vil ligge, kaldes et **95% konfidensinterval for middelværdien**.

Mere præcist gælder det, at hvis man for et stort antal stikprøver på den samme stokastiske variabel angav 95% konfidensintervaller, så ville den sande middelværdi tilhøre 95% af disse intervaller.[1]

[1] **Præcis definition af konfidensinterval.** Lad være givet en stikprøve for en stokastisk variabel *X*, lad β være et tal mellem 0 og 1. Lad endvidere Θ være en punktestimator for parameteren θ og lad *L* og *U* være stokastiske variable, for hvilke det gælder, at $P(L \le \theta \le U) = \beta$. På basis af den givne stikprøve findes tal *l* og *u* som bestemmer det ønskede interval $l \le \theta \le u$. Dette kaldes et $100 \cdot \beta$ procent konfidensinterval for den ukendte parameter θ.

5.3.2. Populationens spredning 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 farvede områder på figuren, 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 $P(\bar{X} \le \bar{x})=0.025$.

Man må her huske, at et gennemsnit har spredningen $\frac{\sigma}{\sqrt{n}}$, hvor σ er spredningen på den enkelte måling og *n* er antal målinger i stikprøven.
Fremfor at løse ovenstående ligning, er det lettere at benytte formlen i sætning 5.1

Sætning 5.1. Spredning kendt eksakt
Er spredningen eksakt kendt er et 95% konfidensinterval bestemt ved formlen

$$\bar{x}-z_{0.975}\cdot\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x}+z_{0.975}\cdot\frac{\sigma}{\sqrt{n}}$$

Bevis:

Af formlen $x_p=\mu+z_p\cdot\sigma$ fås, idet $\sigma(\bar{X})=\frac{\sigma}{\sqrt{n}}$, at $\bar{x}=\mu+z_{0.975}\cdot\frac{\sigma}{\sqrt{n}} \Leftrightarrow \bar{x}-\mu=z_{0.975}\cdot\frac{\sigma}{\sqrt{n}}$

Vi har følgelig, at radius i konfidensintervallet er $r=z_{0.975}\cdot\frac{\sigma}{\sqrt{n}}$ ◆

Sædvanligvis udtrykkes de generelle formler ved signifikansniveauet α, som er sandsynligheden for at begå en fejl. α sættes sædvanligvis til 10%, 5%, 1 % eller 0.1% svarende til henholdsvis 90%, 95%, 99% og 99.9% konfidensintervaller.

I så fald bliver formlen (udtrykt ved α) $\bar{x}-z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x}+z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}$

Alle de anvendte regnemidler har programmer, der automatisk beregner enten konfidensinterval eller radius *r* i konfidensintervallet.

Eksempel 5.2. Konfidensinterval hvis spredningen er kendt eksakt
Lad gennemsnittet af 12 målinger være $\bar{x}=90$, og lad os antage, at spredningen kendes eksakt til $\sigma=0.5$.
Bestem et 95% konfidensinterval for middelværdien μ.
Løsning:
På værktøjslinien foroven: Tryk på = eller f_x ▶Vælg kategorien "Statistisk"▶Vælg "konfidens.norm " ▶udfylde menuen▶
Resultat : radius = 0.283

A	B
KONFIDENS.NORM(0,05;0,5;12)=	0,282896

95% konfidensinterval: [90-0.283;90+0.283] = [89.717 ; 90.283]

Vi ved derfor med 95% "sikkerhed", at populationens sande middelværdi ligger indenfor disse intervaller.
Mere præcist, at af de 100 stikprøver med tilhørende 95% konfidensintervaller, vil i middel kun 5 af disse intervaller ikke indeholde den sande værdi.

5.3.3. Populationens spredning ikke kendt eksakt

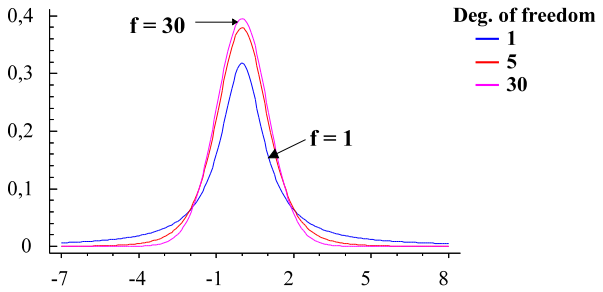
Sædvanligvis er populationens spredning σ jo ikke eksakt kendt, men man regner et estimat s ud for den.
Da s jo også varierer fra stikprøve til stikprøve, giver dette en ekstra usikkerhed, så konfidensintervallet for μ bliver bredere.

Man må så i formlen $\bar{x} - z_{0.975} \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{0.975} \cdot \frac{\sigma}{\sqrt{n}}$ erstatte Z-fraktilen $z_{0.975}$ med en såkaldt T-fraktil $t_{0.975}(f)$ (også benævnt $t_{0.975,f}$) hvor frihedsgradstallet $f = n - 1$, og n = antal målinger).

Mere generelt udtrykt ved signifikansniveauet α erstattes i formlen $\bar{x} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ z- fraktilen $z_{1-\frac{\alpha}{2}}$ med t - fraktilen $t_{1-\frac{\alpha}{2},f}$.

t-fordelinger

En t - fordeling har samme klokkeformede udseende som en Z - fordeling (en normalfordeling med middelværdi 0 og spredning 1)
I modsætning til Z - fordelingen afhænger dens udseende imidlertid af antallet n af tal i stikprøven.
Er **frihedsgradstallet** $f = n$ -1 stort (over 30) er forskellen mellem en U- fordeling og en t-fordeling så lille, at den sædvanligvis er uden praktisk betydning.
Er f lille bliver t - fordelingen så meget bredere end Z - fordelingen, at t-fordelingen må anvendes i stedet for Z-fordelingen.
Grafen viser tæthedsf



Eksempel 5.3. Beregning af t-værdier.

1) Find $t_{0.975}(12)$ og $t_{0.025}(12)$.
2) Find $P(X \ge 1)$, hvor X er t - fordelt med 12 frihedsgrader.

Løsning:
På værktøjslinien foroven: Tryk på f_x ▶Vælg kategorien "Statistisk"▶Vælg "T.INV"
Der fremkommer en tabel med anvisning på, hvordan den skal udfyldes.
1)

A	B
T.INV(0,975;12)	2,178813
T.INV(0,025;12)	-2,17881

$t_{0.975}(12) =$ = 2,178813 $t_{0.025}(12) =$ = - 2,178813

2)

1-T.FORDELING(1;12;1)=	0,168525

$P(X \ge 1) =$

Er den eksakte værdi af spredningen ukendt, erstattes den i den af sætning 5.1 angivne formel med spredningen s.
Idet s har frihedsgradstallet n - 1 erstattes $z_{0.975}$ med $t_{0.975}(n-1)$
Herved fremkommer sætningen.

Sætning 5.2. ***(Den eksakte værdi af spredning ukendt)***

Et 95 % konfidensinterval er bestemt ved formlen: $\bar{x} - t_{0.975}(n-1)\cdot\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{0.975}(n-1)\cdot\frac{s}{\sqrt{n}}$

(eller udtrykt ved α $\bar{x} - t_{1-\frac{\alpha}{2},n-1}\cdot\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{1-\frac{\alpha}{2},n-1}\cdot\frac{s}{\sqrt{n}}$

Eksempel 5.4. Konfidensinterval, hvis spredningen ikke er kendt eksakt.
Ved fremstilling af et bestemt levnedsmiddel er det vigtigt, at et tilsætningsstof findes i levnedsmidlet i en koncentration på 8.50 (g/l).
For at kontrollere dette udtager levnedsmiddelkontrollen 6 prøver af levnedsmidlet. Resultaterne var:

Måling nr	1	2	3	4	5	6
koncentration x (g/l)	8.54	7.89	8.50	8.21	8.15	8.32

Idet man antager, på baggrund af tidligere lignende målinger, at resultaterne er normalfordelte, skal man besvare følgende spørgsmål:.

a) Angiv et estimat for koncentrationens middelværdi og spredning.
b) Angiv et 95% konfidensinterval for koncentrationen, og vurder herudfra om kravet på 8.50 er opfyldt.

Løsning
Excel har indbygget et program, så man ikke behøver at anvende formlerne direkte.
Data indtastes i cellerne A1 til A6 ▶Data ▶ Dataanalyse▶ Beskrivende statistik▶udfyld inputområde▶ vælg Resumestatistik og konfidensniveau

A	B	C	D
8,54		*Kolonne1*	
7,89			
8,5		Middelværdi	8,268333333
8,21		Standardfejl	0,098434976
8,15		Median	8,265
8,32		Tilstand	#I/T
		Standardafvigelse	0,241115463
		Stikprøvevarians	0,058136667
		Kurtosis	-0,2376446
		Skævhed	-0,500530903
		Område	0,65
		Minimum	7,89
		Maksimum	8,54
		Sum	49,61
		Antal	6
		Konfidensniveau(95,0%	0,253035161

a) Resultater: $\bar{x} = 8.268$ og $s = 0.241$.

b) 95% konfidensinterval: $\bar{x} \pm r = 8.268 \pm r$ hvor r = 0.253
[8.268 -0.253 ; 8.268 + 0.253] = [8.02 ; 8.52]
Da intervallet indeholder 8.50, er kravet opfyldt, men da intervallet kun lige netop indeholder tallet 8.50, så det vil nok være rimeligt, at foretage en ny vurdering på basis af nogle flere målinger.

Eksempel 5.5 Konfidensinterval, hvis originale data ikke kendt
Find konfidensintervallet for middelværdien μ, idet stikprøven er på 20 tal, som har et gennemsnit på 50 og en spredning på 12.
Løsning:
Har intet færdigt program, så her må man anvende formlen for konfidensinterval
I kolonne D er de formler angivet, som er brugt i kolonne E
Bemærk, at for overskuelighedens skyld er udskrevet gitterlinier og søjle/række overskrifter

A	B	C	D	E
Eksempel 5.5		**Konfidensradius r =**	**T.INV(1-B6/2;B3-1)*B5/KVROD(B3)=**	**5,616173**
		nedre grænse=	**B4-E1**	**44,38383**
n=	**20**	**Øvre grænse=**	**B4+E1**	**55,61617**
gennemsnit =	**50**			
Spredning=	**12**			
α=	**0,05**			

95% konfidensinterval: [44.38 ; 55.62]

Prædistinationsinterval. Ved mange anvendelser ønsker man at **forudsige**, hvor værdien af en **kommende** observation af den variable med 95%"sikkerhed" vil falde, snarere end at give et 95% konfidensinterval for middelværdien af den variable. Man siger, at man ønsker at bestemme et 95% prædistinationsinterval (forudsigelsesinterval).

SÆTNING 5.2 ($100 \cdot (1-\alpha)$ *% prædistinationsinterval for en enkelt observation)*.
Et $100 \cdot (1-\alpha)$ *% prædistinationsinterval for en enkelt fremtidig observation* X_{n+1} *er bestemt ved*

$$\bar{x} - t_{1-\frac{\alpha}{2}}(n-1) \cdot s \cdot \sqrt{1+\frac{1}{n}} \le \mu \le \bar{x} + t_{1-\frac{\alpha}{2}}(n-1) \cdot s \cdot \sqrt{1+\frac{1}{n}}.$$

Bevis: Lad X_{n+1} være en enkelt fremtidig observation. Eftersom X_{n+1} er uafhængig af de øvrige X'er, er X_{n+1} også uafhængig af $\bar{X}$.
Variansen af differensen $\bar{X} - X_{n+1}$ er følgelig $V(\bar{X} - X_{n+1}) = V(\bar{X}) + V(X_{n+1}) = \frac{\sigma^2}{n} + \sigma^2 = \sigma^2\left(1+\frac{1}{n}\right)$.

Da man sædvanligvis først regner konfidensintervallet ud, så er den nemmeste måde at beregne det tilsvarende prædistinationsinterval at benytte, at radius r_p i prædistinationsinterval fås af radius r_k i konfidensintervallet ved formlen $\boxed{r_p = r_k \cdot \sqrt{1+n}}$

Bevis: $r_p = t_{1-\frac{\alpha}{2}}(n-1) \cdot s \cdot \sqrt{1+\frac{1}{n}} = t_{1-\frac{\alpha}{2}}(n-1) \cdot s \cdot \sqrt{\frac{n+1}{n}} = t_{1-\frac{\alpha}{2}}(n-1) \cdot \frac{s}{\sqrt{n}} \cdot \sqrt{1+n} = r_k \cdot \sqrt{1+n}$

Eksempel 5.6. Prædistinations-interval for middelværdi af normalfordeling.
Samme problem som i eksempel 5.4, men nu ønskes bestemt et 95% prædistinationsinterval for en enkelt ny måling af koncentrationen.
Løsning
Da konfidensintervallet har længden 8.52 - 8.02 = 0.50 er radius $r_k = 0.25$
Vi har derfor $r_p = 0.25 \cdot \sqrt{6+1} = 0.66$ og dermed
95% prædistinationsinterval $= [8.27 - 0.66; 8.27 + 0.66] == \underline{\underline{[7.61 ; 8.93]}}$. ♦

Bestemmelse af stikprøvens størrelse

Før man starter sine målinger, kunne det være nyttigt på forhånd at vide nogenlunde hvor mange målinger man skal foretage, for at få resultat med en given nøjagtighed.
Hvis spredningen antages kendt , ved vi, at radius i konfidensintevallet er

$$r = z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Løses denne ligning med hensyn til *n* fås

$$n = \left(\frac{z_{1-\frac{\alpha}{2}} \cdot \sigma}{r} \right)^2$$

Det grundlæggende problem er her, at man næppe kender spredningen eksakt.
Man kender muligvis på basis af tidligere erfaringer størrelsesordenen af spredningen. Hvis ikke må man eventuelt lave nogle få målinger, og beregne et *s* på basis heraf.
Som en første tilnærmelse antages, at antallet af gentagelser *n* er over 30, så man kan bruge U-fordelingen.
Hvis det derved viser sig, at *n* er under 30 anvendes i stedet en t-fordeling, idet vi løser ligningen

$$n = \left(\frac{t_{1-\frac{\alpha}{2}}(n-1) \cdot \sigma}{r} \right)^2$$

Det følgende eksempel illustrerer fremgangsmåden.

Eksempel 5.7. Bestemmelse af stikprøvens størrelse.
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)
På basis af målingerne på de 7 træer sættes $s \approx 14$.

a) Find hvor mange træer der skal måles, hvis et 95% konfidensinterval højst skal have en radius på ca. 5 cm.
b) Find hvor mange træer der skal måles, hvis et 95% konfidensinterval højst skal have en radius på ca. 6 cm.

Løsning:

a) Først benyttes formlen $n = \left(\frac{z_{0,975} \cdot s}{r} \right)^2$

A	B
(NORM.INV(0,975;0;1)*14/5)^2 =	30,11704

Da $n > 30$ er det rimeligt, at benytte en Z- fordeling frem for en t-fordeling.
Der skal altså tilfældigt udvælges ca. 31 egetræer.

b) Benyttes samme formel som under spm. a) fås n = 21

Da $n < 30$ burde man have anvendt en t - fordeling. $n = \left(\frac{t_{0,975,}(n-1) \cdot s}{r} \right)^2$

Formlen omskrives til $\left(\frac{t_{0,975,}(n-1) \cdot s}{r} \right)^2 - n = 0$

I celle D1 skrives en startværdi for n eksempelvis 21.
I celle F1 skrives= (TINV(0,05;D1)*14/6)^2-D1
Data ▶ Hvad-hvis analyse ▶ "Målsøgning
I "Angiv celle" skrives F1. I "Til Værdi" skrives 0. "Ved ændring af celle" skrives D1

Resultat:

D	E	F
23,41628	(T.INV(0,975;D1-1)*14/6)^2-D1	0

Facit :23,4162
Der skal altså tilfældigt udvælges ca. 24 egetræer.

Da overslaget jo er afhængigt af om vurderingen af s er korrekt, bør man dels for en sikkerheds skyld vælge s lidt rigelig stor, dels efter at man har målt de 31/24 træer lige kontrollere beregningen af konfidensintervallet. ♦

5.4 KONFIDENSINTERVAL FOR SPREDNING

I visse situationer ønsker man at finde et konfidensinterval for spredningen.
Vi vil ikke gå nærmere ind på teorien herfor, men blot henvise til formlerne i oversigt 5.5.
Er den eksakte værdi af middelværdien ukendt skal følgende formel benyttes:

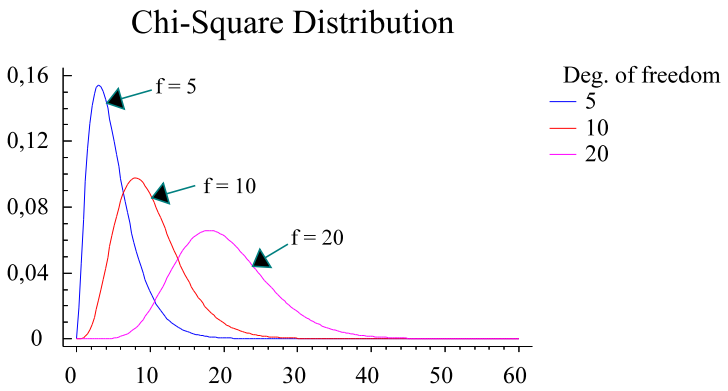
$$\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}$$

I formlerne indgår den såkaldte χ^2 - fordeling, (udtales ki i anden) .

χ^2-fordelinger

χ^2-fordelingen benyttes ved beregninger omkring varianser, når disse er erstattet af et estimat s^2.[1]

På figuren er afbildet tæthedsfunktionen for χ^2-fordelingerne $\chi^2(5)$, $\chi^2(10)$ og $\chi^2(20)$.

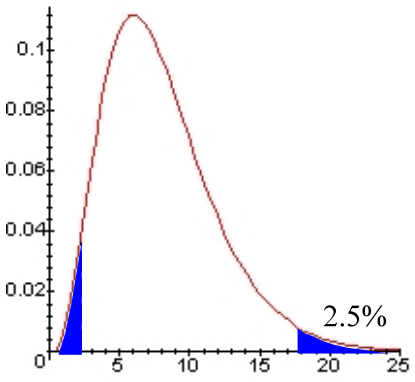


Det ses, at χ^2 kun er defineret for tal større end eller lig nul, og at χ^2-fordelinger ikke er symmetriske om middelværdien. Jo større frihedsgradstallet bliver jo mere symmetriske bliver de dog, og for store f-værdier - i praksis $f > 30$ - kan en χ^2-fordeling $\chi^2(f)$ approksimeres med normalfordelingen $n(\mu,\sigma)$, hvor $\mu = f$ og $\sigma = \sqrt{2 \cdot f}$.

Excel har en kumuleret χ^2-fordeling ligesom naturligvis alle statistikprogrammer har det.

Eksempel 5.8. Beregning af χ^2-værdier.

1) Find $\chi^2_{0.025}(8)$ og $\chi^2_{0.975}(8)$.
2) Find $P(X \le 5)$,
hvor X er χ^2-fordelt med 8 frihedsgrader.



Løsning:

På værktøjslinien foroven: Tryk på f_x ▶Vælg kategorien "Statistisk"▶Vælg "CHI2.inv"
Der fremkommer en tabel med anvisning på, hvordan den skal udfyldes.

1)

CHI2.INV(0,025;8)	2,179731
CHI2.INV(0,975;8)	17,53455

$\chi^2_{0.025}(8) = 2.18$ $\chi^2_{0.975}(8) = 17.5$

2)

CHI2.FORDELING(5;8;1)	0,242424

$P(X \le 5) = \underline{0.242}$

♦

[1] **Definition af χ^2-fordelingen.** *Lad $U_1, U_2, \ldots, U_f$ være uafhængige normerede normalfor-delte variable. Sandsynlighedsfordelingen for den stokastiske variabel $\chi^2 = U_1^2 + U_2^2 + \ldots + U_f^2$ kaldes χ^2-fordelingen med frihedsgradstallet f og betegnes $\chi^2(f)$*

Eksempel 5.9. Konfidensinterval for varians og spredning af normalfordeling.

En virksomhed ønsker at kontrollere med hvilken spredning en bestemt målemetode angiver saltindholdet i en opløsning. Der foretages følgende 12 målinger af en opløsning af det pågældende salt. Resultaterne var:

Måling nr	1	2	3	4	5	6	7	8	9	10	11	12
% opløsning	6.8	6.0	6.4	6.6	6.8	6.1	6.4	6.3	6.0	6.2	5.8	6.2

a) Angiv på basis af måleresultaterne et estimat for opløsningens spredning.

b) Angiv et 95% konfidensinterval for variansen og for spredningen.

Løsning:

Excel har intet færdigt program, så der må anvendes formel 3 i oversigt 5.5 :

$$\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}$$

	A	B	C	D	E
1	6,8	spm. A	s=	STDAFV(A1:A12)	0,316228
2	6				
3	6,4	spm b			
4	6,6	Konfidensinterval for varians			
5	6,8	Nedre grænse		(12-1)*E1^2/CHIINV(0,025;12-1)	0,050182
6	6,1	Øvre grænse		(12-1)*E1^2/CHIINV(0,975;12-1)	0,288279
7	6,4			**[0.0502 ;0.288]**	
8	6,3	Konfidensinterval for spredning			
9	6	Nedre grænse		KVROD(E5)	0,224014
10	6,2	Øvre grænse		KVROD(E6)	0,536916
11	5,8			**[0.224 ; 0.537]**	

a) $\sigma \approx 0.3162$

b) 95% Konfidensinterval for varians [0.05018; 0.28828]
95% konfidensinterval for spredning [0.2240; 0.5369]

5.5. OVERSIGT over centrale formler i kapitel 5

X antages **normalfordelt** $n(\mu,\sigma)$ .Givet stikprøve af størrelsen n med gennemsnit $\bar{x}$ og spredning s

Oversigt over konfidensintervaller

nr	Forudsætninger	Estimat for parameter	100 (1 - α) % konfidensinterval for parameter
1	μ ukendt. σ ukendt	For μ : $\bar{X}$	$\bar{x} - t_{1-\frac{\alpha}{2}}(n-1)\cdot\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{1-\frac{\alpha}{2}}(n-1)\cdot\frac{s}{\sqrt{n}}$ se eksempel 5.4
	μ ukendt. σ kendt	For μ : $\bar{X}$	$\bar{x} - z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}$ se eksempel 5.2
3	μ ukendt σ ukendt.	For σ^2 : s^2	$\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}$
4	μ kendt σ ukendt.	For σ^2 : $s_\mu^2 = \frac{(n-1)s^2 + n(\bar{x}-\mu)^2}{n}$	$\frac{(n-1)s^2 + n(\bar{x}-\mu)^2}{\chi^2_{1-\frac{\alpha}{2}}(n)} \le \sigma^2 \le \frac{(n-1)s^2 + n(\bar{x}-\mu)^2}{\chi^2_{\frac{\alpha}{2}}(n)}$

Oversigt over prædistinationsintervaller

nr	Forudsætninger	Estimat for parameter	100 (1 - α) % konfidensinterval for parameter
1	μ ukendt. σ kendt	For μ : $\bar{X}$	radius i konfidensinterval $r_k = z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}$ radius i prædistinationsinterval $r_p = r_k\sqrt{1+n}$
2	μ ukendt. σ ukendt	For μ : $\bar{X}$	radius i konfidensinterval $r_k = t_{1-\frac{\alpha}{2}}(n-1)\cdot\frac{s}{\sqrt{n}}$ radius i prædistinationsinterval $r_p = r_k\sqrt{1+n}$

Bestemmelse af stikprøvens størrelse *n*.

	Ønsket værdi af radius *r* i 100 (1 - α) % konfidensinterval		
1	σ kendt eller n > 30	$n = \left(\frac{z_{1-\frac{\alpha}{2}}\cdot\sigma}{r}\right)^2$	(norminv(1-$\frac{\alpha}{2}$,0,1)*s/r)^2
2	σ ukendt, men antag den højst er s	$n = \left(\frac{t_{1-\frac{\alpha}{2}}(n-1)\cdot s}{r}\right)^2$	Løs ligning , se eksempel 5.8

OPGAVER

Opgave 5.1
Lad der være givet 10 uafhængige observationer af en syres koncentration (i %).

12.4	10.8	12.1	12.0	13.2	12.6	11.5	11.9	12.8	12.0

1) Find et estimat for koncentrationens middelværdi μ og spredning σ.
2) Angiv et 95% konfidensinterval for μ.
3) Angiv et 95% prædistinationsinterval for en enkelt ny måling af koncentrationen..

Opgave 5.2
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.
3) Angiv et 95% prædistinationsinterval for en enkelt måling af trykstyrken på en ny betonklods.
4) Man fandt, at radius i konfidensintervallet var for stor.
   Bestem med tilnærmelse antallet af målinger der skal udføres, hvis radius højst skal være 12.

Opgave 5.3
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.

Opgave 5.4
En polymer produceres i batch. Viskositetsmålinger udført på hver batch gennem et stykke tid har vist, at variationen i processen er meget stabil med spredning $\sigma = 20$.
På 15 batch gav viskositetsmålingerne følgende resultater:

724	718	776	760	745	759	795	756	742	740	761	749	739	747	742

1) Find et estimat for viskositetens spredning σ.
2) Angiv et 95% konfidensinterval for σ, for at kontrollere påstanden om, at $\sigma = 20$.
3) Find et estimat for viskositetens middelværdi μ.
4) Angiv et 95% konfidensinterval for μ.

Opgave 5.5
Ved en fabrikation af et bestemt sprængstof er det vigtigt, at en reaktoropløsning har en pH-værdi omkring 8.0. Der foretages 6 målinger på en bestemt reaktantopløsning.
Resultaterne var:

pH	8.42	7.36	8.04	7.71	7.65	7.82

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.
3) Man finder, at radius i konfidensintervallet er for bredt.
   Angiv med tilnærmelse antallet af målinger der skal foretages, hvis radius skal være 0.1.

Opgave 5.6
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

a) Angiv et estimat for middelværdien af papirets vægt, og et 95%-konfidensinterval herfor.
b) Angiv med tilnærmelse antallet af ark, der skal anvendes, hvis radius i konfidensintervallet højst skal være r = 0.01
c) Angiv et 95%-prædistinationsinterval for en enkelt nyt ark papir.
d) Angiv et estimat for spredningen og et 95%-konfidensinterval for spredningen af papirets vægt.

Opgave 5.7
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.
2) Opstil et 95% konfidensinterval for målemetodens spredning.

6 HYPOTESETEST (ÉN NORMALFORDELT VARIABEL)

6.1 GRUNDLÆGGENDE BEGREBER

Ofte vil man se vendinger som" Stikprøven viser, at udbyttet ved den ny metode er **signifikant** større end ved den hidtidig anvendte metode"

Statistiske problemer, hvor man på basis af en stikprøve ønsker med eksempelvis 95% "sikkerhed" at bevise en påstand om hele populationen kaldes hypotesetest.

De forskellige begreber der indgår i en hypotesetest vil blive gennemgået i forbindelse med følgende eksempel.

Eksempel 6.1. Hypotesetest.
En fabrik har gennem mange år benyttet en metode, der på basis af en given mængde råmateriale gav et middeludbytte af et produceret stof på μ_0 = 69.2 kg og spredningen σ = 1.0 kg.
En nyansat ingeniør får til opgave at søge at forøge middeludbyttet ved en passende (billig) modifikation af procesbetingelserne.
Efter en række lovende eksperimenter i laboratoriet synes opgaven at være lykkedes, men det endelige bevis herfor er, ud fra et passende antal driftsforsøg statistisk at kunne "bevise", at middeludbyttet er blevet forøget.
Ud fra kendskab til de forskellige mulige støjfaktorer antages spredningen at være uændret på 1.0 kg.
Da driftsforsøgene er meget ressourcekrævende, bevilges der kun 12 delforsøg.
Der foretages 12 uafhængige delforsøg og udbyttet x måltes:

Forsøg nr	1	2	3	4	5	6	7	8	9	10	11	12
x	68.8	70.7	70.3	70.1	70.7	68.7	69.2	68.9	70.0	69.6	71.0	69.1

1) Kan man ud fra disse data bevise på signifikansniveau α = 0.05 , at middeludbyttet er blevet forøget ?

2) Hvis svaret i spørgsmål 1 er bekræftende, så angiv et estimat for det nye middeludbytte, og angiv et 95% konfidensinterval herfor.

Løsning:

1) Løsningen opdeles for overskuelighedens skyld i en række trin

1a) **Definition af stokastisk variabel *X*.**

X = udbyttet ved den modificerede proces.

1b) **Valg af *X*'s fordelingstype.**

X antages at være approksimativt normalfordelt $n(\mu\ , 1.0)$.

1c) **Opstilling af nulhypotese og alternativ hypotese**

Der opstilles en såkaldt **Nulhypotesen H_0 : μ = 69.2 kg.**

Nulhypotesen **skal** indeholde en konkret påstand (her et lighedstegn). Påstanden er, at modifikationen ingen (nul) virkning har

Der opstilles endvidere en **alternativ hypotese *H*: μ > 69.2 kg.**

Den alternative hypotese skal så vidt muligt indeholde det, der ønskes bevist. I dette tilfælde ønskes vist, at middeludbyttet er vokset, dvs. μ > 69.2 kg.

Testen kaldes en **ensidet test** i modsætning til en tosidet test :

H_0 : μ = 69.2 kg contra H: $\mu \neq$ 69.2 kg,

hvor vi blot ønsker at vise, at middeludbyttet har ændret sig.

1d) **Angivelse af testens signifikansniveau.**

Hvis stikprøvens gennemsnit $\bar{x}$ er meget større end 69.2 kg (måske helt op mod 100 kg), så er der stor sandsynlighed for at udbyttet er steget. Man siger så, at **nulhypotesen forkastes**, eller at $\bar{x}$ ligger i forkastelsesområdet (se figur 6.1).

Hvis derimod $\bar{x}$ kun ligger lidt over 69.2 kg, så kan det skyldes tilfældige udsving, og man kan ikke med nogen stor sikkerhed konkludere, at udbyttet er steget. Man siger, at **nulhypotesen accepteres,** eller at $\bar{x}$ ligger i acceptområdet.

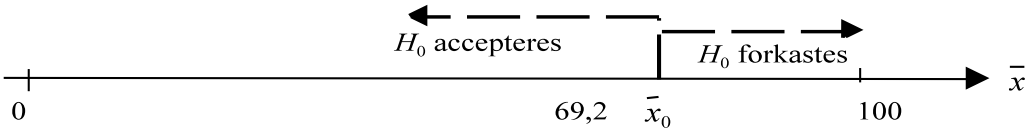


Fig. 6.1 *Accept- og forkastelsesområde*

Lad $\bar{x}_0$ være grænsen mellem acceptområdet og forkastelsesområdet. $\bar{x}_0$ skal bestemmes sådan, at **forudsat H_0 : μ = 69.2 kg er sand**, så er det yderst usandsynligt, at en stikprøves gennemsnit $\bar{x}$ vil komme til at ligge i forkastelsesområdet. Hvis stikprøvens gennemsnit alligevel ligger i forkastelsesområdet, må det være forudsætningen H_0 der er forkert, d.v.s. middeludbyttet må være blevet større.

Det er naturligvis ikke entydigt bestemt, hvad det vil sige, at noget er yderst usandsynligt.

Man starter derfor enhver test med at fastlægge det såkaldte **signifikansniveau** α .

Er α valgt til 5% ,så har man derved fastlagt, at **sandsynligheden for fejlagtigt at påstå,** at middeludbyttet er steget, er under 5%.

Da det kan have alvorlige økonomiske konsekvenser fejlagtigt at påstå at middeludbyttet er steget (produktionen omstilles osv.) ,så er man naturligvis interesseret i, at dette ikke sker.

Det normale i industriel produktion er, at sætte α = 5%, men er det eksempelvis medicinske forsøg, hvor det kan have alvorlige menneskelige konsekvenser, sættes α måske så lavt som 1% eller 0.1%, mens man i andre situationer måske sætter signifikansniveauet til 10%.

I dette eksempel er α sat til 5%.

1e) **Beregning af *P* - værdi**

Gennemsnittet af de 12 resultater giver

A
68,8
70,7
70,3
70,1
70,7
68,7
69,2
68,9
70
69,6
71
69,1

Vælg f_x og MIDDELV og marker listen A

$\overline{x}$ = 69.758 kg.

Under forudsætning af at nulhypotesen H_0 : μ = 69.2 kg er sand, så er $\overline{X}$ er normalfordelt med middelværdi μ_0 = 69.2 og spredning $\frac{\sigma}{\sqrt{n}} = \frac{1.0}{\sqrt{12}} = 0.2887$.

Vi kan derfor nemt finde den præcise adskillelse mellem accept og forkastelsesområdet, da den jo er bestemt ved at arealet skal være 95%

Norm.Inv(0,95;69,2;0,2887)= 69.67

Da 69.76 > 69.76 ligger det målte gennemsnit altså i forkastelsesområdet.

Imidlertid vælger man i stedet at beregne den såkaldte P-værdi (Probability value) som er sandsynligheden for at få en værdi på det fundne stikprøvegennemsnit 69.76 eller derover, dvs. P-værdi = $P(\overline{X} \geq 69.76)$

Er denne P-værdi er mindre end α =0.05 må $\overline{x}$ = 69.76 ligge i forkastelsesområdet (se figur 6.2)

Hvis P-værdien ligger over α ligger $\overline{x}$ = 69.76 i acceptområdet, dvs. vi kan ikke bevise at middeludbyttet er steget.

P-værdi = 1-NORMFORDELING(69,76;69,2;1/KVROD(12);1) = 0,026196

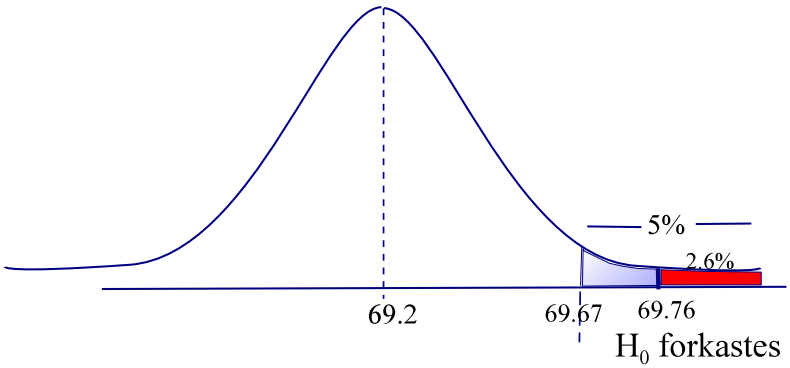


Fig 6.2 *P-værdi*

1f) **Konklusion**
Da P - værdi = 2.62% < 5% forkastes H_0 ,
Vi har et statistisk bevis for, at den modificerede proces giver et større middeludbytte.

2) Udbyttet kan i middel forventes at være ca. $\bar{x} = 69.76$ kg

95% konfidensinterval:
KONFIDENS.NORM(0,05;1;;12) Resultat radius r = 0.5658
C Int :[69.76-0.5658 ; 69,76+0.5658]= [69.19;70.32]

At konfidensintervallet indeholder tallet 69.2 er klart i modstrid med at vi lige har vist, at middelværdien er større end 69.2.
Det skyldes, at konfidensintervallet forkaster med 2.5% til hver side, mens en ensidet test forkaster kun til en side med 5%.
Mere logisk ville det være, at lave en ensidet 95% konfidensinterval,

$$\left[\bar{x} - z_{0.95} \cdot \frac{\sigma}{\sqrt{n}}; \infty\right] = \left[69.76 - 1.65 \cdot \frac{1.0}{\sqrt{12}}; \infty\right] = [69.28; \infty]$$

Det er imidlertid ikke standard, nok fordi det er sværere at forklare en udenforstående, at middelværdien med 95% sikkerhed ligger over 69.28 .

Eksempel 6.2. Hypotesetest, hvor man får accept af H_0.
Samme problem som i eksempel 6.1, men nu er signifikansniveauet α=1%.
Løsning:
H_0: μ = 69.2 mod H: H_0: μ > 69.2
I eksemplet fandt vi på basis af 12 forsøg, at P-værdi = 2.6%.
Konklusion: H_0 accepteres , dvs.
vi kan **ikke** på et signifikansniveau på 1% bevise, at middelværdien var steget.

Bemærk: Vi skriver **ikke** at vi har bevist den ikke er steget, det kan meget vel være tilfældet. Vi kan bare ikke bevise det med den ønskede sikkerhed.

6.2 HYPOTESETEST MED UKENDT MIDDELVÆRDI OG SPREDNING

I eksempel 6.1 blev baggrunden for testen gennemgået. Samtidig antog vi, at spredningen var kendt eksakt. Dette er sjældent tilfældet, men havde vi haft over 30 målinger i stikprøven, ville det sædvanligvis være tilladeligt, at erstatte den eksakte værdi med den beregnede spredning s, og foretage de samme beregninger
Imidlertid er det på et statistisk program lige så let at lave den korrekte t-test, så det vil man nok altid gøre.

Eksempel 6.3. Ensidet hypotesetest om middelværdi (spredning ikke kendt eksakt)

Samme problem som i eksempel 6.1, men nu er spredningen ikke kendt eksakt.

Forsøg nr	1	2	3	4	5	6	7	8	9	10	11	12
x	68.8	70.7	70.3	70.1	70.7	68.7	69.2	68.9	70.0	69.6	71.0	69.1

1) Kan man ud fra disse data bevise på signifikansniveau $\alpha = 0.05$, at middeludbyttet er blevet forøget, dvs. større end 69.2 kg ?
2) Hvis svaret i spørgsmål 1 er bekræftende, så angiv et estimat for det nye middeludbytte, og angiv et 95% konfidensinterval herfor.

Løsning:

1) X = udbyttet ved den modificerede proces.
X antages at være approksimativt normalfordelt $n(\mu\,,\,\sigma)$.
H_0 : μ = 69.2 kg. H: μ > 69.2 kg.

Løsning

Her benyttes formlen i oversigt 6.4.

$P(T \geq t)$, hvor $t = \frac{(\bar{x} - \mu_0)\cdot\sqrt{n}}{s}$ og T er t-fordelt med n -1 frihedsgrader

Data indtastes i A1 til A12

	A	B	C	D	E
1	68,8	x streg =		MIDDEL(A1:A12)	69,75833
2	70,7	s=		STDAFV(A1:A12)	0,816265
3	70,3	Ho		μ0=	69,2
4	70,1	H		μ>μ0	
5	70,7	t=		(E1-E3)*KVROD(12)/E2	2,369481
6	68,7				
7	69,2	P-værdi=		1-T.FORDELING(E5;11;1)	0,018593
8	68,9				
9	70				
10	69,6				
11	71				
12	69,1				

Græske bogstaver findes: Indsæt▶Symbol
Havde der været mere end 30 i stikprøven kunne man tillade sig at bruge Z-test

P-værdi: P($\bar{X}$ >69.2) = 0.0186 =1.86%.
Da P-værdi < 5% forkastes H_0 , dvs. vi har et statistisk bevis for, at den modificerede proces giver et større middeludbytte.

2) Data ▶ Dataanalyse ▶ Beskrivende statistik ▶ udfyld inputområde ▶ vælg konfidensniveau
Resultat : Konfidensniveau(95,0%) radius = 0,51863
Konfidensinterval [69.758-0.517;69.758+0.5179] = [69.24 ; 70.28]

Eksempel 6.4 Tosidet hypotesetest om middelværdi (spredning ikke kendt eksakt).
Ved fremstilling af et bestemt levnedsmiddel er det vigtigt, at et tilsætningsstof findes i levnedsmidler i en koncentration på 8.40 (g/l).
For at kontrollere om tilsætningsstoffet har en koncentration på ca. 8.40, udtager levnedsmiddelkontrollen 6 prøver af levnedsmidler. Resultaterne var:

Måling nr	1	2	3	4	5	6	7	8
Koncentration x (g/l)	8.54	7.89	8.50	8.21	8.15	8.32	8.45	8.31

Det ønskes på denne baggrund undersøgt om koncentrationen har den ønskede værdi.
Signifikansniveau sættes til 5%.
Løsning:
Lad X være koncentrationen af tilsætningsstoffet i levnedsmidlet.
Det antages, at X er normalfordelt $n(\mu, \sigma)$
Da det både er uønsket, at koncentrationen er for lille og at den er for stor, bliver nulhypotesen H_0: μ = 8.4 mod H: $\mu \neq 8.4$, dvs. vi har en tosidet test.

Bemærk, at selv om man vel egentlig hellere ville bevise, at koncentrationen er 8.4 og derfor helst ville have denne påstand i den alternative hypotese, er dette ikke muligt, da nulhypotesen skal indeholde et lighedstegn.
Benytter formler i oversigt 6.4, og beregningerne foregår derfor som i eksempel 6.3.

	A	B	C	D
1	8,54	gennemsr	MIDDEL(A1:A8)	8,29625
2	7,89	spredning	STDAFV.S(A1:A8)	0,21353738
3	8,5	H0	μ0=	8,4
4	8,21	H	μ≠μ0	
5	8,15		(D1-D3)*KVROD(8)/D2	-1,37422923
6	8,32		T.FORDELING.2T(ABS(D5);7,1)	0,21175307
7	8,45			
8	8,31			

Her får vi P-værdi til 0.2117 . Da vi har valgt T.Fordeling.2T er det en tosidet test, dvs. vi får en accept da p-værdi > 0.05
Bemærk, at TI-.. har multipliceret P-værdi med 2, hvilket nok skyldes, at så skal vi altid sammenligne med 5%.
I de tilfælde, hvor man har en tosidet test, kunne man i stedet beregne et konfidensinterval, hvilket er lettere i Excel's tilfælde.

Eksempel 6.5. Test af spredning

En fabrikant af læskedrikke har købt en automatisk "påfyldningsmaskine".
Ved købet af maskinen har man betinget sig, at rumfanget af den påfyldte væske i middel skal have en spredning, der ikke overstiger 0.20 ml.
Efter kort tids anvendelse får man mistanke om, at spredningen er for stor. Mange klager over underfyldte flasker.
Derfor foretages en kontrol, hvor man tilfældigt udtager 20 flasker med læskedrik, og måler rumfanget af væsken i flasken. Det viser sig, at stikprøvens spredning er $s = 0.24$ ml.
Med et signifikansniveau på 5% er det da et statistisk bevis for, at den nye maskine ikke opfylder det stillede krav?

Løsning:

Lad X = rumfang af drik i flaske.
X antages normalfordelt $n(\mu, \sigma)$, hvor såvel μ som σ er ukendte.
H_o: $\sigma = 0.2$ imod H: $\sigma > 0.2$,

$\chi^2 = \frac{(n-1)\cdot s^2}{\sigma_0^2}$ (se oversigt 6.4) dvs. i det foreliggende tilfælde $\chi^2 = \frac{(20-1)\cdot 0.24^2}{0.2^2} = 27.36$.

	A	B
1	(20-1)*0,24^2/0,2^2	27,36
2	1-CHI2.FORDELING(b1;19;1)	0,096543

Da P-værdi=9.65% > 5 %, accepteres H_0, dvs. det er ikke påvist, at spredningen ved påfyldningen er for stor, men der er dog nær ved at være signifikans. ♦

6.3. FEJL AF TYPE I OG TYPE II:

Ved enhver test kan der være to typer fejl, hvoraf vi hidtil kun har taget hensyn til den ene type. For bedre at forstå problemstillingen vil vi se på følgende skema.

		Beslutning	
		H_0 accepteres	H_0 forkastes
Forudsætning	H_0 er sand	Rigtig beslutning	Forkert beslutning Type I fejl
	H_0 er falsk	Forkert beslutning Type II fejl	Rigtig beslutning

Det må være et krav til en god test, at der kun er en lille sandsynlighed for at begå en fejl af type I eller type II.
I eksempel 6.1 ville en type I fejl være, hvis man konkluderer, at den modificerede proces giver et større udbytte, selv om det ikke er tilfældet. Virksomheden bruger måske millionbeløb på at omlægge produktionen, og det er ganske forgæves.
En type II fejl ville være, at man ikke opdager, at den modificerede proces giver et større udbytte. Dette er naturligvis uheldigt, men hvis det skyldes, at forbedringen ikke blev opdaget, fordi den er ganske ringe, har det muligvis ingen praktisk betydning.
Hvis en test har signifikansniveau α og den beregnede P-værdi < α så **forkastes H_0.**
Vi ved hermed, at P(type I fejl) $\leq \alpha$, dvs. vi rimelig sikre på, at have foretaget en korrekt

beslutning.
P-værdien angiver jo nogenlunde sandsynligheden for at vi træffer en forkert beslutning.

Hvis α = 5% og P-værdien er 4.25% forkastes H_0. Det samme sker, hvis P-værdi = 0.001%, men vi er her unægtelig noget sikrere på, at vi at vi træffer en korrekt beslutning.

Hvis vi **accepterer H_0** er det blot udtryk for, at vi ikke kan forkaste(svag konklusion: "H_0 frikendes på grund af bevisets stilling").
Man kan have begået en type II fejl, dvs. ikke opdaget, at den alternative hypotese var sand.

Eksempel 6.6. Fejl af type 2

Samme problem som i eksempel 6.1, men nu er signifikansniveauet α=1%

Løsning:

H_0: μ = 69.2 mod H: H_0: μ > 69.2
I eksemplet fandt vi på basis af 12 forsøg, at P-værdi = 2.6%.
Konklusion: H_0 accepteres , dvs.
vi kan **ikke** på et signifikansniveau på 1% bevise, at middelværdien var steget.
Imidlertid kan middeludbyttet meget vel være steget, men vi kunne bare ikke bevise det med den ønskede sikkerhed. Vi kan have begået en fejl af type 2.

Som det ses af eksempel 6.6, så vil en formindskelse af muligheden for at begå en type 1 fejl (α formindskes) forøge sandsynligheden for at begå en type 2 fejl.
Den eneste måde hvorpå begge kan formindskes er at øge antallet *n* af forsøg.
Problemet hermed er, at man derved måske opdager en så lille forbedring, at det ikke er rentabelt at foretage en dyr ændring af fremstillingsprocessen.
Først når udbyttet overstiger en **bagatelgrænse** Δ vil man reagere.

Dimensionering af forsøg (vælge stikprøvestørrelse *n*).

Lad os antage, at virksomheden i eksempel 6.1 finder, at hvis stigningen i udbyttet ved den modificerede proces er mindre end Δ = 0.5 kg, så har det ingen praktisk interesse (Δ = 0.5 kg er bagatelgrænsen), og derfor gør det intet, hvis man ikke opdager det (begår en type II fejl).
Hvis derimod stigningen Δ er større end 0.5 kg, så har det stor betydning, og sandsynligheden for at begå en type II fejl må derfor være lille. Lad os sætte den til højst β = 10%.

Problemet er nu, hvor stor en stikprøvestørrelse *n* (antallet af delforsøg) der skal udføres, for at ovennævnte krav er opfyldt.

En sådan vurdering kaldes en **dimensionering** af forsøget. Udfører man det ud fra en dimensionering nødvendige antal forsøg, vil en accept af nulhypotesen nu betyde, at nok kan udbyttet være steget, men ikke så meget, at det har praktisk interesse.
I oversigt 6.4 er angivet de formler, der skal anvendes ved en dimensionering.
De følgende 2 eksempler viser anvendelsen heraf.

Eksempel 6.7. Dimensionering (kendt spredning).
Inden man i eksempel 6.1 begyndte at lave de dyre delforsøg, vil ingeniøren gerne have en vurdering af, hvor mange driftsforsøg der er nødvendige, når det vides, at det først er økonomisk rentabelt at gå over til den nye metode, hvis middeludbyttet er steget med mindst 0.5 kg.

1) Find stikprøvestørrelsen n, i det tilfælde, hvor $\Delta = 0.5$ kg og $\beta = 10\%$.
Det antages stadig, at $\sigma = 1.0$ kg og signifikansniveauet er $\alpha = 5$ %.

Lad n være den i spørgsmål 1 fundne stikprøvestørrelse.

2) Idet der udføres n delforsøg skal man besvare følgende spørgsmål:
a) Hvilken konklusion kan drages, hvis man finder, at $\bar{x} = 69.8$
b) Hvilken konklusion kan drages, hvis man finder, at $\bar{x} = 69.4$

Løsning

1) X = udbyttet ved den modificerede proces.
X antages at være approksimativt normalfordelt $n(\mu, 1.0)$.
H_0 : $\mu = 69.2$ kg. H: $\mu > 69.2$ kg.

Da testen er ensidet fremgår det af oversigt 6.4) at: $n \geq \left(\frac{z_{1-\alpha} + z_{1-\beta}}{\frac{\Delta}{\sigma}} \right)^2 = \left(\frac{z_{0.95} + z_{0.90}}{\frac{0.5}{1.0}} \right)^2$

((NORM.INV(0,95;0;1)+NORM.INV(0,9;0;1))/(0,5/1))^2 = 34,25539 n = 35

2a) H_0: $\mu = 69.2$ mod H: H_0: $\mu > 69.2$

P-værdi= 1-NORMAL.FORDELING(69,8;69,2;1/(35)^0,5;1) = 0,000193

Da P-værdi < 0.05 forkastes H_0: $\mu = 69.2$ kg , dvs. vi er på et Signifikansniveau på 5% sikre på at middelværdien er over 69.2 kg.
Imidlertid kan vi ikke være sikre på at den er over bagatelgrænsen 69.2 + 0.5 = 69.7 kg
Lad H_0: $\mu = 69.7$ mod H: H_0: $\mu > 69.7$
Vi finder på samme måde som ovenfor, at P-værdi = 27.7%, dvs. en påstand om at middeludbyttet ligger over 69.7 kg vil være fejlagtig i ca. 28% af tilfældene.
Vi vil derfor næppe på den baggrund gå over til den nye metode.

2b) H_0: $\mu = 69.2$ mod H: H_0: $\mu > 69.2$
Vi finder på samme måde som i punkt 2a) , at P-værdi = 11.8%%,
H_0: $\mu = 69.2$ kg accepteres, dvs. vi kan ikke vise, at middeludbyttet er steget.
Dette kan dog godt være tilfældet, men da vi har dimensioneret er vi rimeligt sikre på, at en eventuel stigning ikke har praktisk interesse. ◆

Eksempel 6.8. Dimensionering, (ukendt spredning)
En virksomhed bliver af miljøkontrollen pålagt at formindske indholdet i sit spildevand af et stof A, der mistænkes for at kunne forurene grundvandet. Indholdet af stoffet A i spildevandet skal under 1.7 mg/l, og miljøkontrollen henviser til en ny metode, som burde kunne formindske indholdet til det ønskede niveau. For at vurdere den nye metode ønskes foretaget en række delforsøg.
Hvor mange forsøg skal der mindst foretages, hvis $\alpha = 5\%$, $\beta = 10\%$, $\Delta = 0.10$ mg/l og et overslag over hvor stor σ er sætter denne til 0.15 mg/l.

Løsning:
Lad X = indhold af A (i mg/l) efter benyttelse af den ny metode.
X antages normalfordelt $n(\mu, \sigma)$, hvor såvel μ som σ er ukendte.

Da indholdet af stoffet A ønskes formindsket, bliver
nulhypotesen $H_0 : \mu = 1.7$ mg/l mod $H: \mu < 1.7$ mg/l, dvs. vi har en ensidet test.
Da σ ikke er kendt (kun et løst skøn kendes), er testen en t - test.
Formlen i oversigt 6.4 anvendes:

Først beregnes $n \geq \left(\frac{z_{0.95} + z_{0.90}}{\frac{\Delta}{\sigma}} \right)^2$

((NORMINV(0,95;0;1)+NORMINV(0,9;0;1))/(0,1/0,15))^2 Resultat n = 19.27
Da n < 30 bør man nu løse en ligning (se nedenfor)
Da spredningen jo var usikker, så vil man nok nøjes med at sætte n = 30

Præcis beregning: Løs ligningen $n = 19.27 \cdot \left(\frac{t_{0.95}(n-1)}{z_{0.95}} \right)^2$

Resultatet 19.27 anbringes i celle A1
I celle B1 skrives som startværdi for n tallet 19 .
▶ I celle C1 skrives =A1*(TINV(0,10;B1-1)/NORMINV(0,95;0;1))^2-B1 ▶
Data ▶ Hvad-hvis analyse ▶ "Målsøgning I "Angiv celle" skrives C1. I "Til Værdi" skrives 0. "Ved ændring af celle" skrives B1
Resultat: I celle B1står 21,18523 dvs. n = 22

<u>Den ønskede dimensionering kræver altså 22 forsøg.</u>

6.4. OVERSIGT over centrale formler i kapitel 6

X antages **normalfordelt** $n(\mu,\sigma)$.Givet stikprøve af størrelsen n med gennemsnit $\bar{x}$ og spredning s
Signifikansniveau: α. μ_0 er en given konstant

Oversigt over test af middelværdi μ

T er en stokastisk variabel der er t - fordelt med $f = n - 1$.

Y er en stokastisk variabel, der er normalfordelt $n(\mu_0, \frac{\sigma}{\sqrt{n}})$

Forudsætninger	Alternativ hypotese H	P - værdi	Beregning	H_0 forkastes
σ **ukendt.** $t = \frac{(\bar{x}-\mu_0)\cdot\sqrt{n}}{s}$.	$H: \mu > \mu_0$	$P(T \geq t)$	se eksempel 6.3	P - værdi $< \alpha$
	$H: \mu < \mu_0$	$P(T \leq t)$	se eksempel 6.3, idet man dog ikke kun har t.fordeling	
	$H: \mu \neq \mu_0$	$P(T \geq t)$ for $\bar{x} > \mu_0$ $P(T \leq t)$ for $\bar{x} \leq \mu_0$	som række 1 som række 2	P - værdi $< \frac{1}{2}\alpha$
σ **kendt eksakt**	$H: \mu > \mu_0$	$P(Y \geq \bar{x})$	1-Normfordeling($\bar{x}$;μ_0;1/ *Kvrod*(n);1)	P - værdi $< \alpha$
	$H: \mu < \mu_0$	$P(Y \leq \bar{x})$	Normfordeling($\bar{x}$;μ_0;1/ *Kvrod*(n);1)	
	$H: \mu \neq \mu_0$	$P(Y \geq \bar{x})$ for $\bar{x} > \mu_0$ $P(Y \leq \bar{x})$ for $\bar{x} \leq \mu_0$	som række 1 som række 2	P - værdi $< \frac{1}{2}\alpha$

Dimensionering

$\Delta = |\mu - \mu_0|$ er den mindste ændring i μ der har praktisk interesse. α =P(type I fejl), β = P(type II fejl)

Forudsætning	Hypotese	Formel	Beregning
σ **kendt eksakt**	Ensidet	$n \geq \left(\frac{z_{1-\alpha}+z_{1-\beta}}{\frac{\Delta}{\sigma}}\right)^2$	((NORMINV(1- α ;0;1)+NORMINV(1- β;0;1))/(Δ / σ))^2
	Tosidet	$n \geq \left(\frac{z_{1-\frac{\alpha}{2}}+z_{1-\beta}}{\frac{\Delta}{\sigma}}\right)^2$	((NORMINV((1- α)/2;0;1)+NORMINV(1- β;0;1))/(Δ / σ))^2
σ er ukendt, men erstattes i formlerne af det bedste estimat eller gæt for spredningen.	Ensidet	$n \geq \left(\frac{z_{1-\alpha}+z_{1-\beta}}{\frac{\Delta}{\sigma}}\right)^2 \cdot \left(\frac{t_{1-\alpha}(n-1)}{z_{1-\alpha}}\right)^2$	Løse ligning, se eksempel 6.10
	Tosidet	$n \geq \left(\frac{z_{1-\frac{\alpha}{2}}+z_{1-\beta}}{\frac{\Delta}{\sigma}}\right)^2 \cdot \left(\frac{t_{1-\frac{\alpha}{2}}(n-1)}{z_{1-\frac{\alpha}{2}}}\right)^2$	Løse ligning, se eksempel 6.10

Oversigt over test af varians σ^2

Q er χ^2 fordelt med $f = n - 1$.
σ_0 er en given konstant

Forudsætning	Alternativ hypotese H	*P* - værdi	Beregning	H_0 forkastes
μ **ukendt** $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$	$H: \sigma^2 > \sigma_0^2$	$P(Q \geq \chi^2)$	1-CHI2.Fordeling(χ^2;n-1;1)	*P*-værdi< α
	$H: \sigma^2 < \sigma_0^2$	$P(Q \leq \chi^2)$	CHI2.Fordeling(χ^2;n-1;1)	
	$H: \sigma^2 \neq \sigma_0^2$	$P(Q \geq \chi^2)$ for $\chi^2 \geq n-1$ $P(Q \leq \chi^2)$ for $\chi^2 < n-1$	som række 1 som række 2	*P*-værdi< $\frac{1}{2}\alpha$
μ **kendt** $\chi^2 = \frac{(n-1)s^2 + n(\bar{x}-\mu)^2}{\sigma_0^2}$	$H: \sigma^2 > \sigma_0^2$	$P(Q \geq \chi^2)$	1-CHI2.Fordeling(χ^2;n;1)	*P*-værdi< α
	$H: \sigma^2 < \sigma_0^2$	$P(Q \leq \chi^2)$	CHI2.Fordeling(χ^2;n;1)	
	$H: \sigma^2 \neq \sigma_0^2$	$P(Q \geq \chi^2)$ for $\chi^2 \geq n-1$ $P(Q \leq \chi^2)$ for $\chi^2 < n-1$	som række 1 som række 2	*P*-værdi< $\frac{1}{2}\alpha$

OPGAVER

Opgave 6.1
Et levnedsmiddel ("corned beef") forhandles i pakker på 100 g.
Ved fabrikationen tilsættes traditionelt et konserveringsmiddel B (nitrit).
Da man har mistanke om, at B anvendt i større mængder kan have uønskede bivirkninger, må der højst tilsættes 2.5 mg B pr. 100 g.
Fabrikanten reklamerer med, at der i middel højst er 2 mg B pr. pakke.
En konkurrent tvivler herpå, og vil teste påstanden.
Der købes i forskellige butikker i alt 36 pakker, og indholdet af B blev målt.
Man fandt et gennemsnit af B på $\bar{x}$ = 2.10 mg med et estimat på spredningen på s = 0.30 mg .
Kan man ud fra disse data bevise på signifikansniveau α = 0.01, at reklamen lyver.

Opgave 6.2
Et flyselskab overvejer at lukke en flyrute, såfremt μ = "middelværdien af antal solgte pladser pr. afgang" er under 60.
På de sidste n = 100 afgange er der i gennemsnit solgt $\bar{x}$ = 58.0 pladser med en standardafvigelse på s =11.0 pladser.

1) Kan man ud fra disse data bevise på signifikansniveau α = 0.05, at der i middel er solgt under 60 pladser pr. afgang? (Husk at anføre: Hvad X er. Antagelser. Nulhypotese. Beregninger. Konklusion).
2) Forudsat, at man i spørgsmål 1 kan bevise, at der er solgt under 60 pladser, skal der angives et estimat $\tilde{\mu}$ for middelværdien μ samt et 95% konfidensinterval for middelværdien.

Opgave 6.3
En fabrikation er baseret på en kemisk reaktion, hvor processen forudsætter tilstedeværelse af en katalysator. Med den hidtil benyttede katalysatortype C_1 udnyttes i middel kun ca. 70% af den dyreste råvare. Firmaet overvejer at gå over til en mere effektiv katalysatortype C_2 ved produktionen. Omlægning hertil vil imidlertid kræve betydelige etableringsomkostninger, hvorfor firmaet kun vil lægge produktionen om, såfremt i middel mindst 80% af den dyreste råvare udnyttes, når C_2 benyttes. Til vurdering heraf foretoges en række forsøg med benyttelse af C_2.
Følgende udnyttelsesprocenter fandtes:

68.3	87.7	80.0	84.2	84.0	83.6	76.4	79.9	89.3	75.8
96.1	88.0	79.8	83.7	84.4	95.5	84.2	92.1	92.4	83.9

1) Vurder, om de opnåede forsøgsresultater kan opfattes som et eksperimentelt bevis for, at i middel over 80% af den dyreste råvare udnyttes, når C_2 benyttes. α = 1%
2) Forudsat, at man i spørgsmål 1 kan bevise, at over 80% af den dyreste råvare udnyttes, skal der angives et estimat $\tilde{\mu}$ samt et 95% konfidensinterval for middelværdien μ .

Vi antager i det følgende, at udnyttelsesprocenten X (approksimativt) er normalfordelt n ($\bar{x}$,s)

3) Beregn sandsynligheden for, at udnyttelsesprocenten X for en enkelt måling er mindre end 80%, når C_2 benyttes.

Opgave 6.4

Et kemikalium fremstilles industrielt ved inddampning af en bestemt opløsning. Det var vigtigt, at denne opløsning var svagt basisk med pH = 8.0. Man foretog derfor kontrolmæssigt nogle pH-bestemmelser for den benyttede opløsning. Følgende værdier fandtes:

8.2	8.3	7.9	8.2	7.8	8.6	8.9	7.8	8.2

a) Foretag en testning af om opløsningen kan antages at opfylde kravet til pH-værdi
b) Forudsat, at man i spørgsmål a) kan bevise, at opløsningen ikke opfylder kravet, skal opstilles et 95% konfidensinterval for pH-værdien.

Opgave 6.5

Man frygter, at den såkaldte " syreregn er årsag til, at en bestemt skov er stærkt medtaget. Man måler SO_2 - koncentrationen forskellige steder i skovbunden (i μ g/m^3) og finder:

32.7	23.9	21.7	18.6	27.6	35.1	42.2	36.5	13.4	41.8	34.3	30.0

I ubeskadede skove er SO_2 - koncentrationen 20 μ g/m^3.

a) Giver forsøgene et bevis for, at middelkoncentrationen af SO_2 i den beskadigede skov er større end normalt?
b) Forudsat, at man i spørgsmål a kan bevise, at middelkoncentrationen af SO_2 i den beskadigede skov er større end normalt, skal man angive et tosidet 95%-konfidensinterval for SO_2 - koncentrationen.

Opgave 6.6

Et nyt måleapparat påstås at give måleresultater med spredningen σ = 1.8 mg/l ved måling af salt-indholdet i en opløsning. Da dette er mindre end det sædvanlige, køber et laboratorium et eksemplar af apparatet for at kontrollere påstanden.
Der foretages 15 målinger med følgende resultater:

3.4	7.7	6.0	8.1	8.4	2.7	4.9	1.2	2.1	5.4	3.5	1.5	5.2	4.1	3.9

Test på basis af disse resultater, om spredningen afviger fra 1.8 mg/l.
(Husk altid at anføre: Hvad X er. Antagelser. Nulhypotese. Beregninger. Konklusion.).

Opgave 6.7

Ved indkøbet af et nyt måleapparat oplystes det, at apparatet målte med en spredning på 2.8 enheder. Efter at have brugt apparatet et stykke tid nærede køberen mistanke om, at apparatet målte med større spredning end oplyst.
For at få spørgsmålet undersøgt lod køberen en bestemt måling udføre et antal gange.
Følgende resultater fandtes:

18.8	15.5	12.2	14.8	4.80	1.20	1.43	9.60	1.39	1.17	5.60	1.27	1.35
8.70	1.23	1.40	1.02	1.65	1.91	1.14	1.46	1.59	1.54	1.01	1.80	

Hvilke konklusioner kan køberen drage ud fra en statistisk analyse af de fundne forsøgsresultater?

Opgave 6.8

På et kraftvarmeværk mener man, at en ny metode vil kunne formindske svovlindholdet i de slagger, der bliver tilbage efter kulfyringen. Med en bestemt kvalitet kul, har det hidtidige svovlindhold været 2.70 %.

For at vurdere den nye metode ønsker ingeniøren at foretage en række forsøg.

1) Hvor mange forsøg skal der mindst foretages, hvis α = 5%, β = 10%, Δ = 0.04 og et overslag over spredningens størrelse sætter den til højst 0.08%.
2) Uanset resultatet af dimensioneringen i spørgsmål 1), er der kun praktiske muligheder for at lave 16 forsøg. Følgende værdier af svovlindholdet fandtes (%).

2.58	2.64	2.80	2.50	2.52	2.69	2.60	2.73	2.61	2.62	2.65	2.58	2.70	2.67	2.62	2.64

   Test om disse måleresultater beviser, at svovlindholdet ved den nye metode i middel er blevet mindre.
3) Er det på basis af resultaterne muligt at vurdere, om den fundne formindskelse er stor nok til, at man vil gå over til den nye metode?

Opgave 6.9

På pakken af en iscreme står, at portionen indeholder 14 gram fedt. For at kontrollere dette købes *n* pakker is, og fedtindholdet måles.

1) Bestem den nødvendige stikprøvestørrelse *n,* for at man ved en forskel i fedtindhold på Δ = 0.40 gram højst har, at *P* (type I fejl) = α = 0.01 og *P* (type II fejl) = β = 0.05. ($\sigma \approx 0.42$ gram).
2) Man finder et gennemsnit på 13.1 gram og et estimat s for spredningen på 0.42 gram. Kan man ud fra disse data bevise på signifikansniveau α = 0.01, at middelindholdet afviger fra 14 gram? (Husk altid at anføre: Hvad *X* er. Antagelser. Nulhypotese. Beregninger. Konklusion.).
3) Er det på basis af resultaterne muligt at vurdere, om at en eventuel afvigelse er større end bagatelgrænsen på 0,4 gram.

7. REGNEREGLER FOR SANDSYNLIGHED, KOMBINATORIK

7.1 REGNEREGLER FOR SANDSYNLIGHEDER

Vi har tidligere omtalt sandsynlighed.
I dette kapitel omtales nogle af de grundlæggende definitioner og begreber
Det følgende eksempel blive benyttet til illustration af definitioner og begreber.

Eksempel 7.1. Gennemgående eksempel.
To skytter Anders og Brian skyder hver ét skud mod en skydeskive. Sandsynligheden for at Anders rammer skiven er 0.80 mens Brian har en træfsandsynlighed på 0,60.
Et eksperiment består i at de hver skyder et skud.
Lad A være hændelsen at Anders rammer skiven og lad B være sandsynligheden for at Brian rammer skiven.
Vi har derfor, at P(A) = 0.80 og P(B) = 0.60. ♦

Lad os ved at sætte en streg over A forstå "ikke A".
Generelt gælder $P(\overline{A})$ = 1 - P (A)
I eksempel 8.1 er $\overline{A}$ hændelsen at Anders ikke rammer skiven.
Vi har derfor, at P($\overline{A}$) = 1 - P(A) = 1 - 0.8 = 0.20

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 8.1).
Eksempelvis er $A \cap B$ i eksempel 7.1 hændelsen, **at både Anders og Brian** rammer skiven

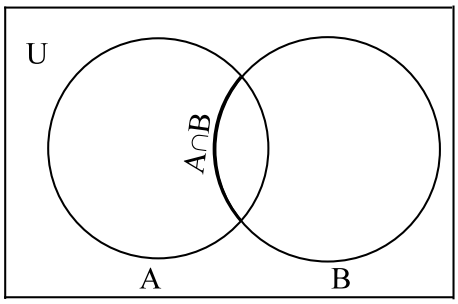


Fig 7.1. *Fællesmængde*

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 7.2)
Eksempelvis er $A \cup B$ i eksempel 8.1 den hændelse, at enten rammer Anders eller også rammer Brian skiven eventuelt gør de det begge.
Man kunne også udtrykke det ved at mindst en af dem rammer skiven.

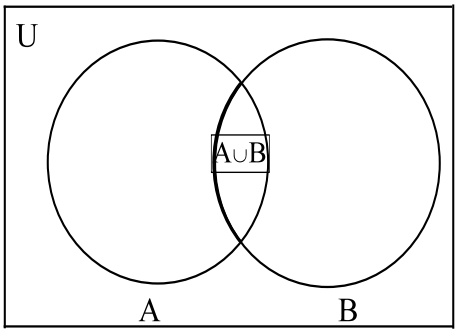


Fig. 7.2 *Foreningsmængde*

Der gælder nu følgende **additionssætning**: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Sætningen fremgår umiddelbart ved at betragte arealerne i figur 7.3.

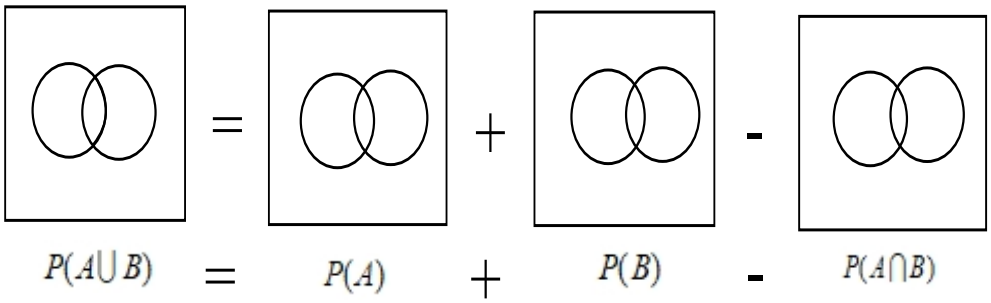


Fig.7.3 *Additionssætning*

Statistisk uafhængighed.
To hændelser A og B siges at være **statistisk uafhængige**, såfremt sandsynligheden for, at den ene hændelse indtræffer, ikke afhænger af, om den anden hændelse indtræffer.
I eksempel 7.1 må man eksempelvis antage, at om Anders rammer skiven har ingen indflydelse på om Brian rammer, så her må man antage A og B er uafhængige.

Et andet eksempel er kast med en terning. Her vil sandsynligheden for at få en sekser i andet kast være uafhængigt af udfaldet i første kast

Der gælder følgende **Produktsætning for uafhængige hændelser:**

For to uafhængige hændelser gælder $P(A \cap B) = P(A) \cdot P(B)$

Eksempel 7.2 (eksempel 7.1 fortsat)
Lad A være hændelsen, at Anders rammer skiven, og lad B være hændelsen, at Brian rammer skiven. Det er givet, at P(A) = 0.80 og P(B) = 0.60.
Find sandsynligheden for
a) At både Anders og Brian rammer skiven
b) At enten Anders eller Brian (evt. begge) rammer skiven, dvs. mindst en af dem rammer skiven.
c) At hverken Anders eller Brian rammer skiven
Løsning:
a) Da hændelserne antages at være uafhængige gælder ifølge produktsætningen $P(A \cap B) = 0.8 \cdot 0.6 = \underline{\underline{0.48}}$
b) Ifølge additionssætningen gælder $P(A \cup B) = 0.6 + 0.8 - 0.48 = \underline{\underline{0.92}}$
c) $P(\overline{A} \cap \overline{B}) = P(\overline{A}) \cdot P(\overline{B}) = (1-0.8)(1-0.6) = \underline{\underline{0.08}}$ ♦

Produktsætning og additionssætning kan generaliseres til flere hændelser end 2.
For tre hændelser A, B og C gælder således
$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C)$
I tilfælde af at hændelserne A, B og C er uafhængige gælder således:
$P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C)$.

Er hændelserne A og B ikke uafhængige, kan man som beskrevet i afsnit 7.3 udlede en mere generel produktsætning

7.2. Betinget sandsynlighed

Er hændelserne A og B ikke uafhængige vil $P(A \cap B) \neq P(A) \cdot P(B)$

Eksempel 7.3. Ikke uafhængige hændelser
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.
Her er $P(A) \cdot P(B) = 0.1 \cdot 0.2 = 0.02 \neq P(A \cap B) = 0.01$.

For at få en mere generel regel indføres $P(B|A)$ 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 7.3, 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

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

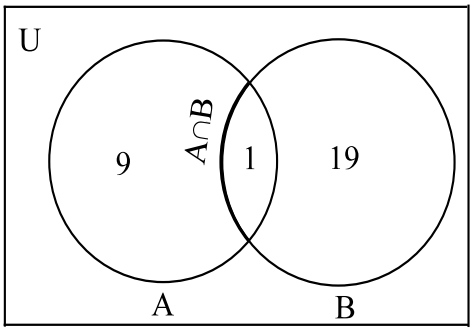


Fig. 7.4 *Taleksempel*

Denne beregning begrunder rimeligheden i følgende definition:
Den af A **betingede sandsynlighed** for B **skrives** $P(B|A)$ (sandsynligheden for, at B indtræffer, når A er indtruffet) defineres ved $P(B|A) = \frac{P(A \cap B)}{P(A)}$.
Ved multiplikation fås **Produktsætningen:** $P(A \cap B) = P(A) \cdot P(B|A)$

Benyttes produktsætningen på eksempel 7.1 fås $P(A \cap B) = P(A) \cdot P(B|A) = 0.1 \cdot 0.1 = 0.01$.

Eksempel 7.4: 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 $P(B|A) = \frac{3}{6}$ og derfor ifølge produktsætningen $P(A \cap B) = P(A) \cdot P(B|A) = \frac{1}{4}$
2) Her er $P(B|A) = \frac{2}{5}$ 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**" : $P(B|A) = \frac{P(B) \cdot P(A|B)}{P(A)}$

Bevis:

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

♦

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

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

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 $P(K) = \frac{16}{100} = 16\%$ og $\underline{\underline{P(T|K) = \frac{4}{16} = 25\%}}$ ♦

7.3. Kombinatorik

7.3.1. 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.

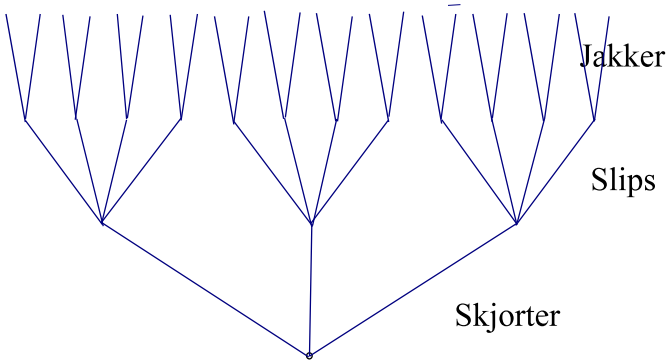
I sådanne tilfælde, bliver problemet derfor, hvorledes man let kan optælle antal udfald. Dette kan ofte gøres ved benyttelse af **kombinatorik**.

7.3.2. 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 7.6. Multiplikationsprincippet
En mand ejer 2 forskellige jakker, 4 slips og 3 forskellige fabrikater skjorter.
På hvor mange forskellige måder kan han sammensætte sin påklædning af skjorte,slips og jakke.
Løsning:
1) Valg af skjorte giver 3 valgmuligheder
2) Valg af slips giver 4 valgmuligheder
3) Valg af jakke giver 2 valgmuligheder
Ifølge multiplikationsprincippet giver det i alt $2 \cdot 3 \cdot 4 = \underline{\underline{24}}$ muligheder
Man kunne illustrere løsningen ved følgende "forgreningsgraf"



Eksempel 7.7 Fakultet
På hvor mange måder kan 5 personer opstilles i en kø (i rækkefølge)
Løsning:
Pladserne i køen nummereres 1,2,3,4,5.
Plads nr. 1 i køen besættes 5 valgmuligheder
Plads nr. 2 i køen besættes 4 valgmuligheder
Plads nr. 3 i køen besættes 3 valgmuligheder
Plads nr. 4 i køen besættes 2 valgmuligheder
Plads nr. 5 i køen besættes 1 valgmulighed
I alt $5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$ forskellige rækkefølger.

Ved ***n* fakultet (*n* udråbstegn) forstås** $n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1$
Endvidere defineres 0! = 1.
f_x ▶Matematik og trigonometri ▶fakultet (5) 5! = 120

7.3.3 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

a) **Uden tilbagelægning**

Eksempel 7.8. 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 er
   $10 \cdot 9 \cdot 8 \cdot 7 = \underline{5040}$
   f_x ▶ **Statistisk** ▶ **PERMUT(10;4)** Resultat: = 5040

♦

Eksempel 7.8 begrunder følgende definition

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

b) **Med tilbagelægning**

Eksempel 7.9. 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 multiplikationsprincippet) $10 \cdot 10 \cdot 10 \cdot 10 = 10^4$ ♦

7.3.4. Uordnet stikprøveudtagelse

Eksempel 7.10 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 kombination af *r* elementer fra M er et udvalg af *r* elementer udtaget af M uden at tage hensyn til rækkefølgen af elementer

Antallet af kombinationer med *r* elementer betegnes K(*n*,*r*) eller $\binom{n}{r}$ (*n* over *r*).

Sætning 7.1 (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 $P(5,3)=K(5,3)\cdot 3! \Leftrightarrow K(5,3)=\frac{P(5,3)}{3!} \Leftrightarrow K(5,3)=\frac{5\cdot 4\cdot 3}{3!}=\frac{5\cdot 4\cdot 3\cdot 2\cdot 1}{3!\,2!}=\frac{5!}{3!\cdot 2!}$ ♦

Eksempel 7.11. 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

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

f_x ▶Matematik og trig ▶KOMBIN(10;4)

♦

OPGAVER

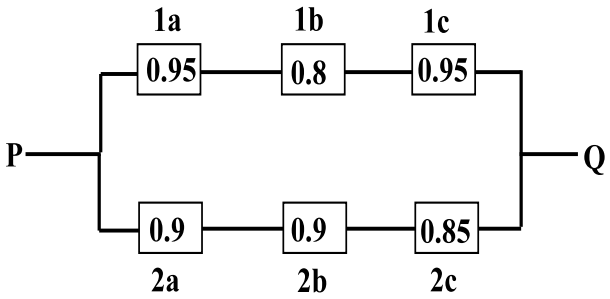
Opgave 7.1

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 følgende hændelser.

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

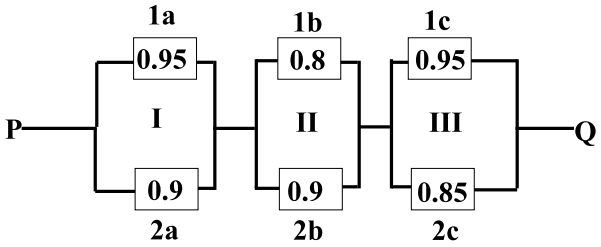
Opgave 7.2

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 7.3
Tre skytter skyder hver ét skud mod en skydeskive. De har træfsandsynligheder 0.75, 0.50 og 0.30.
Beregn sandsynligheden for
1) ingen træffere, 2) én træffer, 3) to træffere, 4) tre træffere.

Opgave 7.4
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 7.5
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 7.6
Af en forsamling på 8 kvinder og 4 mænd skal udtages en arbejdsgruppe på 5 personer.
a) Gør rede for, at gruppen kan udvælges på 448 forskellige måder, når det forlanges, at den skal bestå af højst 3 kvinder og højst 3 mænd.
b) Beregn antallet af måder, hvorpå gruppen kan udvælges, når det forlanges, at de 5 personer ikke alle må være af samme køn.

Opgave 7.7
a) Bestem det antal måder, hvorpå bogstaverne A, B og C kan stilles rækkefølge.
b) Samme opgave for A, B, C og D.

Opgave 7.8.
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 7.9
Bestem antallet af 5-cifrede tal, der kan skrives med to l-taller, et 2- tal og to 3-taller.

Opgave 7.10
Fire projektgrupper på en virksomhed antages at have sandsynlighederne 0.6, 0.7, 0.8 og 0.9 for at få succes med deres projekt. Grupperne antages at arbejde uafhængigt af hinanden. Find sandsynligheden for, at
a) alle grupper får succes,
b) ingen grupper får succes,
c) mindst 1 gruppe får succes,
d) i alt netop 1 gruppe får succes,
e) i alt netop 3 grupper får succes,
f) i alt netop 2 grupper får succes.

Opgave 7.11

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.

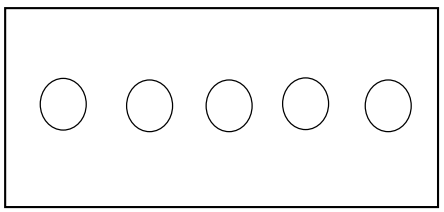
Opgave 7.12

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 7.13

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 7.14

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 7.15

Hvor mange forskellige telefonnumre på 8 cifre kan man danne, når første ciffer ikke må være nul?

8. VIGTIGE DISKRETE FORDELINGER

8.1 INDLEDNING

Vi vil i dette kapitel betragte **diskrete** stokastiske variable, hvis værdier er hele tal.
Vi vil især behandle de diskrete fordelinger:
"Den hypergeometriske fordeling", "Binomialfordelingen" og "Poissonfordelingen"

8.2 HYPERGEOMETRISK FORDELING

Den "hypergeometriske fordeling", finder bl.a. anvendelse ved kvalitetskontrol af varepartier (jævnfør eksempel 8.3), 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 8.1. 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.

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

f_x ▶Hypgeo.fordeling(1;4;5;13;0)=0.3916

Karakteristisk for en hypergeometrisk fordeling er, at elementerne i udfaldsrummet (kugler i en beholder) kan opdeles i **to** grupper.
En opdeling kunne som i eksempel 8.1 være kvinder og mænd eller som i kvalitetskontrol være i defekte varer og ikke-defekte varer.

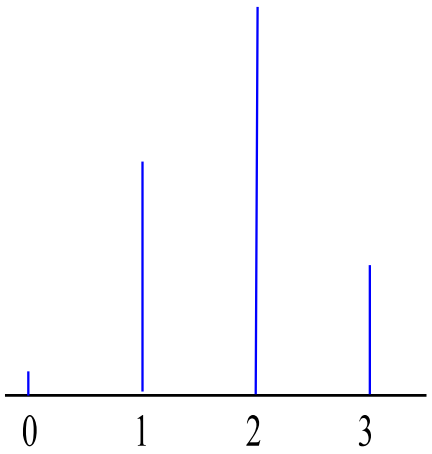
Lad os antage, at vi har en beholder med N kugler, hvoraf de M er røde og resten har en anden farve.
Der udtrækkes en stikprøve på n kugler **uden tilbagelægning**.
Lad X være antallet af røde kugler blandt de n kugler.
X er hypergeometrisk fordelt med parametrene N, M, n (kort skrevet h(N,M,n))
P(X = x) er sandsynligheden for at netop x kugler er røde blandt de n udtrukne kugler.

X siges at være **hypergeometrisk** fordelt med parametrene N, M, n (kort skrevet h(N,M,n)) hvor
Formlen udledes på samme måde som det skete i eksempel 8.1
Sætte x = 0, 1, 2, ... finder vi forskellige værdier af tæthedsfunktionen.

I "Supplement til statistiske grundbegreber" afsnit 8A bevises, at den hypergeometriske fordeling har middelværdien $E(X) = n \cdot p$ og spredningen $\sigma(X) = \sqrt{n \cdot p \cdot (1-p) \cdot \frac{N-n}{N-1}}$, hvor $p = \frac{M}{N}$.

Eksempel 8.2: Hypergeometrisk fordeling *h* (9, 6, 3).
I en urne findes 10 kugler, hvoraf 6 er sorte, 4 er hvide.
Vi betragter det tilfældige eksperiment: "Udtrækning af en kugle og observation af farven på kuglen". Eksperimentet gentages 3 gange, idet den udtrukne kugle ikke lægges tilbage mellem hver udtrækning.
Lad X betegne antallet af udtrukne sorte kugler.
Find og skitser tæthedsfunktionen for X, og beregn middelværdi og spredning for X.
Løsning:
X er en diskret stokastisk variabel, der som er hypergeometrisk fordelt h (10, 6, 3) med tæthedsfunktionen $f(x) = P(X = x)$:

$$f(x) = P(X = x) = \begin{cases} \frac{K(6,0) \cdot K(4,3)}{K(10,3)} = \frac{4}{120} = 0.033 & \text{for} \quad x = 0 \\ \frac{K(6,1) \cdot K(4,2)}{K(10,3)} = \frac{36}{120} = 0.300 & \text{for} \quad x = 1 \\ \frac{K(6,2) \cdot K(4,1)}{K(10,3)} = \frac{60}{120} = 0.500 & \text{for} \quad x = 2 \\ \frac{K(6,3) \cdot K(4,0)}{K(10,3)} = \frac{20}{120} = 0.167 & \text{for} \quad x = 3 \\ 0 & \text{ellers} \end{cases}$$



Stolpediagram for h (10, 6, 3).

Sættes $p = \frac{M}{N} = \frac{6}{10}$ er middelværdien $E(X) = n \cdot p = 3 \cdot \frac{6}{10} = 1.8$ og
spredningen $\sigma(X) = \sqrt{n \cdot p \cdot (1-p) \cdot \frac{N-n}{N-1}} = \sqrt{3 \cdot \frac{6}{10} \cdot \left(1 - \frac{6}{10}\right) \cdot \frac{10-3}{10-1}} = \underline{\underline{0.748}}$

◆

Den hypergeometriske fordeling finder bl.a. anvendelse i kvalitetskontrol, hvilket følgende eksempel viser.

Eksempel 8.3: Stikprøveudtagning (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, såfremt udtrækningen sker **uden** mellemliggende tilbagelægninger ?

Løsning:
X = antal defekte blandt de 25 komponenter
X er hypergeometrisk fordelt med N = 600, M=10, og n = 25
Da partiet godkendes, hvis der **enten** er 0 defekte **eller** 1 defekt, følger af additionssætningen at

$$p = P(X = 0) + P(X = 1).$$

Hændelsen "$X = 0$" forudsætter, at vi i alt udtager 0 af de 10 defekte og 25 forskellige af de 590 ikke-defekte, dvs. $P(X=0) = \dfrac{K(10,0)\cdot K(590,25)}{K(600,25)} = 0.6512$.

Hændelsen "$X = 1$" forudsætter, at vi i alt udtager 1 af de 10 defekte og 24 forskellige af de 590 ikke-defekte, dvs. $P(X=1) = \dfrac{K(10,1)\cdot K(590,24)}{K(600,25)} = 0.2876$.

Vi har altså $p = 0.6512 + 0.2876 = 0.9388 = \underline{93.88\%}$.
P(X ≤ 1)
Vælg f_x ▶Statistik ▶HYPGEOFORDELING ▶Udfyld menu ▶

HYPGEO.FORDELING(1;25;10;600;1)	0,938876

P(X ≤ 1) = 0.9389

Bemærk, at skrives 1 til sidst fås summen, skrives 0 eller intet skrives punktsandsynligheden

8.3 BINOMIALFORDELING

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.

Eksempel 8.4. En binomialfordelt variabel.
En drejebænk producerer 1 % defekte emner.
Lad X være antallet af defekte blandt de næste 5 emner der produceres.
Vi ønsker at finde sandsynligheden for at finde netop 2 defekte blandt disse 5, det vil sige $P(X=2)$.
Løsning:
Lad et eksperiment være at udtage et emne fra produktionen.
Resultatet af eksperimentet har to udfald: defekt, ikke defekt.
Eksperimentet gentages 5 gange uafhængigt af hinanden.
Der er en bestemt sandsynlighed for at få en defekt, nemlig $p = 0.01$.
Lad d være det udfald at få en defekt, og $\overline{d}$ være det udfald at få en fejlfri.
Vi opskriver nu samtlige forløb, der giver 2 defekte ud af 5

$d,d,\bar{d},\bar{d},\bar{d}$
$d,\bar{d},d,\bar{d},\bar{d}$
$d,\bar{d},\bar{d},d,\bar{d}$
$d,\bar{d},\bar{d},\bar{d},d$
$\bar{d},d,d,\bar{d},\bar{d}$
$\bar{d},d,\bar{d},d,\bar{d}$
$\bar{d},d,\bar{d},\bar{d},d$.
$\bar{d},\bar{d},d,d,\bar{d}$
$\bar{d},\bar{d},d,\bar{d},d$
$\bar{d},\bar{d},\bar{d},d,d$

Da eksperimenterne gentages uafhængigt af hinanden, følger det af produktsætningen (både -og), at det første forløb må have sandsynligheden

$$0.01\cdot 0.01\cdot(1-0.01)\cdot(1-0.01)\cdot(1-0.01)=0.01^2\cdot(1-0.01)^3.$$

Det næste forløb må have sandsynligheden

$$0.01\cdot(1-0.01)\cdot 0.01\cdot(1-0.01)\cdot(1-0.01)=0.01^2\cdot(1-0.01)^3$$

Vi ser, at alle gunstige forløb har samme sandsynlighed.
Antal forløb må være lig antal måder man kan placere 2 *d*'er på 5 tomme pladser (eller antal måder man kan tage 2 kugler ud af en mængde på 5).
Dette ved vi kan gøres på K(5,2)=10 måder (svarende til de 10 forløb).
Vi får følgelig, at $p=K(5,2)\cdot 0.01^2\cdot(1-0.01)^3=0.00097$

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
X er en diskret stokastisk variabel med tæthedsfunktionen

$$f(x)=P(X=x)=\begin{cases}K(n,x)\cdot p^x\cdot(1-p)^{n-x} & \text{for} \quad x\in\{0,1,2,\ldots,n\}\\ 0 & \text{ellers}\end{cases}$$

X siges at være binomialfordelt b (n, p).

Eksempel 8.5 = eksempel 8.4 fortsat
X er binomialfordelt b(n,p) hvor n = 5 og p = 0.01
P(X=2) :
Vælg f_x ▶Statistik ▶BINOMIALFORDELING ▶Udfyld menu ▶

BINOMIAL.FORDELING(2;5;0,01;0)	0,00097

P(X=2) = 0.00097

SÆTNING 8.1. (*middelværdi og spredning for binomialfordeling*).
Lad X være binomialfordelt b (*n, p*).
Der gælder da $E(X) = n \cdot p$ og $\sigma(X) = \sqrt{n \cdot p \cdot (1-p)}$.

Bevis:
Lad os betragte et eksperiment, hvor resultatet "succes" har sandsynligheden *p* for at ske.
Lad os foretage *n* **uafhængige gentagelser af eksperimentet**. At gentagelserne er uafhængige betyder, at udfaldet af et eksperiment ikke afhænger af udfaldet af de forrige eksperimenter.
Lad os betragte n stokastiske variable $X_1, X_2, \ldots, X_n$,

hvor $X_i = \begin{cases} 1 & \text{hvis i'te gentagelse af eksperimentet giver succes.} \\ 0 & \text{ellers} \end{cases}$

Vi har $E(X_i) = \sum_i x_i f(x_i) = 1 \cdot p + 0 \cdot (1-p) = p$, og

$$V(X_i) = \sum_i (x_i - \mu)^2 f(x_i) = (1-p)^2 \cdot p + (0-p)^2 \cdot (1-p) = p - p^2 = p \cdot (1-p)$$

Idet $X = X_1 + X_2 + \ldots + X_n$ er binomialfordelt *b* (*n*, *p*) fås af linearitetsreglen (kapitel 1afsnit 5), at
$E(X) = E(X_1) + E(X_2) + E(X_3) + \ldots + E(X_n) = p + p + p + \ldots + p = n \cdot p$.
Endvidere fås af kvadratreglen i kapitel 1 afsnit 5, idet vi har uafhængige gentagelser, at
$V(X) = V(X_1) + V(X_2) + \ldots + V(X_n) = p \cdot (1-p) + p \cdot (1-p) + \ldots + p \cdot (1-p)$,
eller $V(X) = n \cdot p \cdot (1-p)$.

◆

Eksempel 8.6: Tæthedsfunktion for binomialfordelt variabel .
Lad der på to af sidefladerne på en terning være skrevet tallet 1, på to andre sideflader være skrevet tallet 2 og på de sidste to sideflader være skrevet tallet 3.Vi betragter det tilfældige eksperiment:
"7 kast med en terningen og observation af det fremkomne tal.
Lad *X* betegne antallet af toere ved de 7 kast. *X* antages at være binomialfordelt $b\left(7, \frac{1}{3}\right)$.

1) Angiv tæthedsfunktionen *f* (*x*) for *X* (3 betydende cifre), og tegn et stolpediagram for *f* (*x*).
2) Find middelværdi og spredning for *X*

En person foretager eksperimentet 11 gange, d.v.s. foretager 11 gange en serie på 7 kast med terningen. Stikprøven gav følgende resultat

Antal toere i en serie	0	1	2	3	4	5	6	7
Antal gange dette skete	1	2	4	3	1	0	0	0

3) Giv på grundlag af stikprøven et estimat for *p* i binomialfordelingen.
4) Giv på grundlag af stikprøven et estimat for middelværdi og spredning

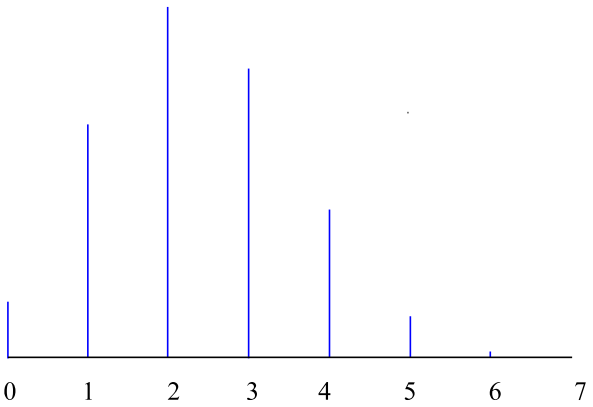
Løsning:

1) f(x) = P(X = x) = K(7,x)$\cdot\left(\frac{1}{3}\right)^{x}\cdot\left(1-\frac{1}{3}\right)^{n-x}$ hvor x = 0, 1, 2 osv.

og derefter x = 1 osv.

BINOMIAL.FORDELING(0;7;1/3;0)	0,058528

$$f(x)=P(X=x)=\begin{cases}0.059 & \text{for} \quad x=0\\ 205 & \text{for} \quad x=1\\ 0.307 & \text{for} \quad x=2\\ 0.256 & \text{for} \quad x=3\\ 0.128 & \text{for} \quad x=4\\ 0.038 & \text{for} \quad x=5\\ 0.006 & \text{for} \quad x=6\\ 0.000 & \text{for} \quad x=7\\ \text{ellers} & 0\end{cases}$$



Stolpediagram for binomialfordelingen

2) $E(X)=n\cdot p=7\cdot\frac{1}{3}=\underline{\underline{2.33}}$ og $\sigma(X)=\sqrt{n\cdot p\cdot(1-p)}=\sqrt{7\cdot\frac{1}{3}\cdot\left(1-\frac{1}{3}\right)}=\underline{\underline{1.25}}$

3) Der er i alt $1\cdot 0+2\cdot 1+4\cdot 2+3\cdot 3+1\cdot 4=23$ toere i 77 kast.

Et estimat for p er $\hat{p}=\frac{23}{77}=\underline{\underline{0.299}}$

4) Stikprøvens middelværdi er $\bar{x}=\frac{23}{11}=\underline{\underline{2.09}}$, og stikprøvens spredning er

$$\sigma(X)=\sqrt{n\cdot p\cdot(1-p)}=\sqrt{7\cdot\frac{23}{77}\cdot\left(1-\frac{23}{77}\right)}=\underline{\underline{1.21}}$$

◆

Approksimation af hypergeometrisk fordeling med binomialfordeling.

At erstatte den hypergeometriske fordeling *h* (*N*, *M*, *n*) med binomialfordelingen *b* (*n*, *p*) vil for de fleste anvendelser kunne gøres med en passende nøjagtighed, hvis stikprøvestørrelsen *n* er mindre end eller lig 10% af partistørrelsen *N* ($n\le\frac{N}{10}\Leftrightarrow\frac{n}{N}\le\frac{1}{10}$).

I så fald sættes i binomialfordelingen $p=\frac{M}{N}$.

Eksempel 8.7. Approksimation af hypergeometrisk fordeling til binomialfordeling.
I eksempel 8.3, hvor man udtog 25 komponenter fra æsker på 600 komponenter, skete udtagningen logisk nok uden tilbagelægning. Imidlertid er det klart, at da æskerne indeholder mange komponenter vil sandsynligheden for at få en defekt ikke ændrer sig meget, hvis man i stedet havde foretaget udtagningen med tilbagelægning.
Der blev antaget, at der var 10 defekte i en sådan æske med 600, og dette antal defekte vil så være konstant, under hver udtrækning.

Vi har derfor, at P(at få en defekt) = $\frac{25}{600} < \frac{1}{10}$

Betingelserne for at benytte binomialfordelingen er nu til stede.
Løsningen af problemet i eksempel 8.3 vil derfor nu være:

$p_a = P(X \le 1) = P(X = 0) + P(X = 1)$ = BINOMIAL.FORDELING(1;25;1/60;1) = 0.9353

Det ses, at vi får praktisk samme resultat som i eksempel 8.3. ◆

Hypotesetest for binomialfordelt variabel.
I kapitel 6 gennemgik vi ved en række eksempler de grundlæggende begreber for hypotesetestning for én normalfordelt variabel. Disse begreber kan uændret overføres til hypotesetestning for binomialfordelt variabel.

Konfidensintervaller.
Ved løsning af en passende ligning (se oversigt 8.8) kan man finde de eksakte grænser for konfidensintervallerne.
Som beskrevet i appendix er det ofte muligt at approksimere med en normalfordeling.
Derved fremkommer de formler som er oversigt 8.8.
Inden de anvendes skal man undersøge om approximationen gælder.

De følgende to eksempler viser beregning af test og konfidensintervaller.

Eksempel 8.8. Ensidet binomialfordelingstest.
En levnedsmiddelproducent fremstiller et levnedsmiddel A, som imidlertid har en ret ringe holdbarhed. Efter en række eksperimenter lykkedes det at frembringe et produkt B, som i alt væsentligt er identisk med A, men som har en bedre holdbarhed. Af markedsmæssige grunde er det vigtigt, at der ikke er forskel på smagen af B og af det velkendte produkt A. For at undersøge dette, lader producenten et panel af 24 ekspertsmagere vurdere, om man kan smage forskel. Man foretog derfor følgende smagsprøvningseksperiment.

Hver ekspertsmager fik 3 ens udseende portioner, hvoraf en portion var af det ene levnedsmiddel og de to andre portioner var af det andet levnedsmiddel.

Hvilket af de 3 portioner der skulle indeholde et andet levnedsmiddel end de to andre, og om det skulle være levnedsmiddel A eller B , afgjordes hver gang ved lodtrækning. Kun forsøgslederen havde kendskab til resultatet.
Hver ekspertsmager fik besked på, at de skulle fortælle forsøgslederen hvilken af de tre portioner der smagte anderledes. Hvis man ikke kunne smage forskel, skulle man gætte.

Resultatet viste, at af de 24 svar var 13 svar rigtige.
Ved ren gætning kunne man forvente ca. $\frac{1}{3}$ dvs. ca. 8 rigtige svar. 13 rigtige svar er betydeligt flere, men kan det alligevel tilskrives tilfældigheder ved gætning?

Kan der på et signifikansniveau på 5% statistisk påvist, at ekspertsmagerne kan smage forskel på smagen af A og B?

Løsning:

Lad X = antallet af rigtige svar.

X er binomialfordelt b (n, p), hvor n = 24 og p er ukendt.

Nulhypotese $H_0: p = \frac{1}{3}$ mod den alternative hypotese $H: p > \frac{1}{3}$

P - værdi = $P(X \geq 13) =$ | P(X>=13)= | 1-BINOMIAL.FORDELING(12;24;1/3;1) | 0,028441 |

Da P - værdi = 2.84% < 5% forkastes nulhypotesen (enstjernet), dvs. der må konkluderes, at der er en smagsforskel mellem produkt A og B. ◆

Eksempel 8.9. Konfidensinterval for parameteren p i binomialfordeling.

En plastikfabrik har udviklet en ny type affaldsbeholdere. Man overvejer at give en 6 års garanti for holdbarheden. For at få et skøn over om det er økonomisk rentabelt, bliver 100 beholdere udsat for et accelereret livstidstest som simulerer 6 års brug af beholderne. Det viste sig, at af de 100 beholdere overlevede de 85 testen.

Idet antallet af overlevende beholdere antages at være binomialfordelt, skal man

1) Angive et estimat for sandsynligheden p for at en beholder "overlever" i 6 år .
2) Angive et 95% konfidensinterval for p.

Løsning:

1) Lad X være antallet af "overlevende" beholdere.
   X forudsættes binomialfordelt b (100, p).

   Ifølge oversigt 8.8 er et estimat for p: $\tilde{p} = \frac{x}{n} = \frac{85}{100} = \underline{\underline{0.85}}$

2) **Eksakt løsning:**

   Benyttes formel i oversigt 8.8.

   I celle A1 skrives en startværdi for p eksempelvis 0,5.►
   I celle B1 skrives =BINOMIAL.FORDELING(85;100;A1;1)►
   Data► What if analyse►"Målsøgning
   I "Angiv celle" skrives B1. I "Til Værdi" skrives 0,025. I "Ved ændring af celle" skrives A1.

A	B	C
0,913554	0,024981	BINOMIAL.FORDELING(85;100;A1;1)

Resultat p = 0.914

Nedre grænse: : Løs ligningen $P(X \leq 85) = 0.975$ med hensyn til p.
Samme metode, men nu skrives 0.975 fremfor 0.025

0,775556	0,975943	BINOMIAL.FORDELING(85;100;A3;1)

Resultat p = 0.776

95% Konfidensinterval: [0.765; 0.914]

Bemærk, at konfidensintervallet ikke ligger helt symmetrisk omkring 0.85, da binomialfordelingen ikke er helt symmetrisk omkring 0.85

Forklaring på formlen:

Udenfor et 95% konfidensinterval ligger 5%, og af symmetrigrunde ligger der 2,5% på hver side.
Er den sande værdi for p eksempelvis 90% vil der i middel være 90 ud af 100 overlevende beholdere. Nu fandt vi kun 85 ud af 100. Sandsynligheden P(X≤85) for at få 85 eller færre overlevende beholdere ud af 100 er derfor ret ringe. Vi har P(X≤85) =binomCdf(100,0.9.0.85) = 0.0726. Selv om 7.26% er et lille tal, så er det dog over

2.5%, så en P-værdi på 90% ligger inde i konfidensintervallet.
For at finde den øvre grænse må vi derfor løse ligningen P(X≤85) = 0.025 med hensyn til p.
Dernæst findes nedre grænse ved at lade p falde, indtil $P(X \geq 85) \approx 0.025$

Approksimativ løsning

Man bruger formlen, der findes i oversigt 8.8: $\tilde{p} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}} \leq p \leq \tilde{p} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}$

r = | NORM.INV(0,975;0;1)*KVROD(0,85*(1-0,85)/100) | 0,069985 |

Resultat: 95% konfidensinterval :[0.85-0.07 ; 0.85+0.07 = [0.78 ; 0.92]
Som det ses, er forskellen i forhold til den eksakte metode meget lille.

◆

Bestemmelse af stikprøvens størrelse

Før man starter sine målinger, kunne det være nyttigt på forhånd at vide nogenlunde hvor mange målinger man skal foretage, for at få resultat med en given nøjagtighed.
Hvis man antager, at man kan approksimere med normalfordelingen, ved vi, at radius for et 95% konfidensinterval er $r = z_{0.975} \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}}$.

Løses denne ligning med hensyn til n fås $n = \left(\frac{z_{0.975}}{r}\right)^2 \hat{p} \cdot (1-\hat{p})$

Det grundlæggende problem er her, at man næppe kender $\hat{p}$ eksakt.

Man kender muligvis på basis af tidligere erfaringer størrelsesordenen af $\hat{p}$. Hvis ikke kunne man eventuelt udtage en lille stikprøve, og beregne et $\hat{p}$ på basis heraf.

Endelig er der den mulighed, at sætter $\hat{p} = 0.5$, som er maksimumsværdien af $\hat{p} \cdot (1-\hat{p})$

Benyttes denne værdi får man den størst mulige værdi af n for en given værdi af r.
Ulempen er, at dette fører til en større stikprøvestørrelse end nødvendigt.
Det følgende eksempel illustrerer fremgangsmåden.

Eksempel 8.10. Bestemmelse af antal i stikprøve.

I en opinionsundersøgelse vil man spørge et repræsentativt antal vælgere om hvilket parti de vilde stemme på, hvis der var valg i morgen.
I denne undersøgelse ønskes inden udtagning af stikprøven, at antallet skal være så stort, at radius i konfidensintervallet højst er 2%.

Løsning:

Metode 1. For at få en øvre grænse, sættes $\hat{p} = 0.5$.

Vi får $n = \left(\frac{z_{0.975}}{r}\right)^2 \hat{p} \cdot (1-\hat{p}) =$ | (NORM.INV(0,975;0;1)/0,02)^2*1/2*1/2 | 2400,912 |

Metode 2 Da man på forhånd ved, at ved sidste valg fik ingen partier mere end 30% af stemmerne sættes $\hat{p} = 0.3$.

$n = \left(\frac{z_{0.975}}{r}\right)^2 \hat{p} \cdot (1-\hat{p}) =$ ORM.INV(0,975;0;1)/0,02)^2*0,3*0,7 | 2016,766

◆

8.4 POISSONFORDELINGEN

Poissonfordelinger benyttes ofte som statistisk model for antallet af "impulser" pr. tidsenhed. Disse impulser antages at komme **tilfældigt og uafhængigt af hinanden**.

Som eksempler kan nævnes: Antal trafikuheld på en bestemt vejstrækning i løbet af et år, antal biler, der passerer en militær kontrolpost, antal varevogne der ankommer pr. time til et stort varehus og antal telefonsamtaler der føres fra en telefoncentral, der er oprettet under en øvelse. Modellen kan dog også anvendes på andet end pr. tidsenhed, eksempelvis også på antal revner pr. km kabel, hvis disse revner forekommer tilfældigt og uafhængigt af hinanden.

Under sådanne omstændigheder kan man ofte benytte den i det følgende omtalte Poissonfordeling som statistisk model for antallet af "impulser" pr. tidsenhed eller volumenenhed eller længdeenhed osv.

SÆTNING 8.2 (*Poissonfordeling*). *Lad X være en stokastisk variabel, som angiver antallet af impulser i et givet tidsrum (eller areal, volumen, produktionsenhed osv.), idet ethvert tidspunkt i tidsrummet har samme mulighed for at være impulstidspunkt som ethvert andet tidspunkt. Endvidere skal impulserne indtræffe tilfældigt og uafhængigt af hinanden* *).
Hvis det gennemsnitlige antal impulser i tidsrummet er $\mu > 0$ *, så siges X at være Poissonfordelt* $p(\mu)$ *med sandsynlighedsfordelingen (tæthedsfunktionen) f(x) = P(X = x) bestemt ved*

$$f(x) = P(X = x) = \begin{cases} \frac{\mu^x}{x!} \cdot e^{-\mu} & \text{for} \quad x \in \{0,1,2,\ldots\} \\ 0 & \text{ellers} \end{cases}$$

Middelværdien for $p(\mu)$ *er* $E(X) = \mu$ *og spredningen er* $\sigma(X) = \sqrt{\mu}$.

I formuleringen af de ovennævnte betingelser kan efter behov "et lille tidsrum Δt*" erstattes med "en lille længde* $\Delta \ell$*", "et lille areal* ΔA*" eller "et lille volumen* ΔV*".*

*) Præcis formulering: Følgende 3 betingelser skal være opfyldt:

1) Sandsynligheden for netop én impuls i et meget lille tidsrum Δt er med tilnærmelse proportional med Δt .

   (Matematisk formulering $\lim_{\Delta t \to 0} \frac{P(X=1)}{\Delta t} = \lambda$ (λ er en positiv konstant)

2) Sandsynligheden for 2 eller flere impulser i det meget lille tidsrum Δt er lille sammenlignet med Δt.

   (Matematisk formulering $\lim_{\Delta t \to 0} \frac{P(X>1)}{\Delta t} = 0$)

3) Antal impulser i forskellige, ikke overlappende tidsrum er statistisk uafhængige.

En bevisskitse for sætningen kan ses i "Supplement til statistiske grundbegreber" afsnit 8.C.

Eksempel 8.11: Antal revner p. meter i et tyndt kobberkabel.

På en fabrik fremstilles kobberkabler af en bestemt tykkelse. Mikroskopiske revner forekommer tilfældigt langs disse kabler. Man har erfaring for, at der i gennemsnit er 12.3 af den type revner p. 10 meter kabel.

Beregn sandsynligheden for, at der

1) ingen revner er i 1 meter tilfældigt udvalgt kabel.
2) er mindst 2 revner i 1 meter tilfældigt udvalgt kabel.
3) er højst 4 revner i 2 meter tilfældigt udvalgt kabel

Fabrikken går nu over til en anden og billigere produktionsmetode. For at få et estimat for middelværdien ved den nye metode måltes antallet af revner på 12 kabelstykker på hver 10 meter. Resultaterne var

Kabel nr.	1	2	3	4	5	6	7	8	9	10	11	12
Antal revner	8	4	14	6	8	10	10	16	2	2	6	8

4) Angiv på basis heraf et estimat for middelværdien af antal revner pr. 10 m kabel.

Løsning:

X = antal revner i 1 meter kabel.

X antages Poissonfordelt $p(\mu)$. (idet vi med tilnærmelse kan antage, at betingelserne i sætning 8.2 er opfyldt (impuls er her ridser).

Da det gennemsnitlige antal revner pr. 1m kabel er $\mu = \frac{12.3}{10} = 1.23$ fås:

1) P(X=0)= POISSON(0;1,23;0) =0,292293 = 0.292
2) $P(X \geq 2) = 1 - P(X \leq 1) =$ 1 - POISSON(1;1,23;1) = 0,348188
3) Y = antal revner i 2 meter kabel.

   Da der i gennemsnit er 2,46 revner i 2 meter kabel, er 2.46 et estimat for μ .

   Vi har derfor $P(X \leq 4)$ =POISSON(4;2,46;1) = 0,896458

4) Der er i alt 94 revner i 12 kabelstykker på hver 10 meter.

   Et estimat for μ er derfor $\tilde{\mu} = \frac{94}{12} = 7.83$. ◆

Hypotesetest og konfidensintervaller for Poissonfordelt variabel.

I kapitel 5 gennemgik vi ved en række eksempler de grundlæggende begreber for hypotesetestning og konfidensintervaller for én normalfordelt variabel.

Disse begreber kan uændret overføres til hypotesetestning og konfidensintervaller for Possonfordelt variabel.

Har man rådighed over et program med kumuleret Poissonfordeling kan testene gennemføres eksakt. (se oversigt 8.8)

Eksempel 8.12. Ensidet Poissontest.

I eksempel 8.11 betragtede vi mikroskopiske revner i et kobberkabel. Fabrikken gik over til en anden og billigere produktionsmetode.

1) Test, om den nye metode giver færre revner end den gamle metode.

2) Forudsat, den nye metode giver signifikant færre revner end den gamle metode, skal man

2a) Angiv et 95% konfidensinterval for middelværdien μ af antal revner pr. 120 meter kabel

2b) Angiv et 95% konfidensinterval for middelværdien μ_1 af antal revner pr. 10 meter kabel.

Løsning:

1) Lad *X* betegne antallet af revner i 120 meter kabel ved ny metode

X antages Poissonfordelt $p(\mu)$, hvor vi i eksempel 8.7 fandt at et estimat for μ var $\tilde{\mu} = 94$.

Ved gammel metode er antal revner i 120 m kabel i middel $\mu_0 = 12.3 \cdot 12 = 147.6$

Nulhypotese $H_0: \mu = 147.6$ mod den alternative hypotese $H: \mu < 147.6$.

P - værdi = $P(Y \le 94)$ = Poisson(94;147,6;1) = 1,52403E-06

Da P - værdi < 0.05 forkastes nulhypotesen (stærkt) ,dvs. vi er sikre på, at middelantallet af revner er blevet formindsket ved at anvende den nye metode

2a) Excel kan ikke regne de exakte grænser ud, men da m = 94 >10 kan approksimeres med normalfordelingen (se oversigt 9.8)

$$m - z_{1-\frac{\alpha}{2}}\sqrt{m} \le \mu \le m + z_{1-\frac{\alpha}{2}}\sqrt{m}$$

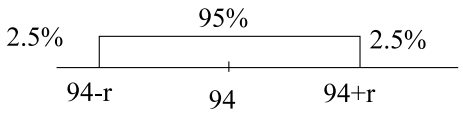
A	B	C	D
m =	94	B1-NORM.INV(0,975;0;1)*KVROD(B1)	74,99744
		B1+NORM.INV(0,975;0;1)*KVROD(B1)	113,0026

95% Konfidensinterval: [75.0; 113.0]

Formlen er indtastet direkte (starte med = og hente Kvrod fra matematik og trigonometri

Forklaring på formlen:

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



Jo mindre den sande værdi af middelværdien er, jo mindre er sandsynligheden for, at gennemsnittet blev 94.

Vi leder derfor i grænsen efter et x , så $P(X \le x) = 0.025$

b) 10 m kabel: $\left[\frac{75}{12}; \frac{113}{12}\right] = [6.25; 9.41]$ ◆

8.5 APPROKSIMATIONER

Vi har undertiden benyttet os af, at det under visse forudsætninger er muligt med en rimelig nøjagtighed, at foretage approksimationer, f.eks. at approksimere en binomialfordeling eller en Poissonfordeling med en normalfordeling.

Dette kan give nogle simplere beregninger, eksempelvis når man approksimerer en hypergeometrisk fordeling med en binomialfordeling eller når man ved udregning af konfidensintervaller for binomialfordeling approksimerer med normalfordeling.

I appendix 8.1 er angivet en samlet oversigt over de mulige approksimationer.

8.6 Den generaliserede hypergeometriske fordeling.

Den hypergeometriske fordeling benyttes som model ved stikprøveudtagning uden tilbagelægning, hvor hvert element har enten en bestemt egenskab (defekt) eller ikke har denne egenskab (ikke defekt). Hvis der foreligger flere end to egenskaber, f.eks. udtagning af møtrikker, hvis diameter enten tilhører et givet toleranceinterval eller er for stor eller for lille, kan man generalisere den hypergeometriske fordeling. Dette illustreres ved følgende eksempel:

Eksempel 8.13. Generaliseret hypergeometrisk fordeling.

I en urne findes 12 kugler, hvoraf 5 er sorte, 4 er hvide og 3 er røde.

Vi betragter det tilfældige eksperiment: "Udtrækning af 6 kugler uden tilbagelægning og observation af farven på kuglerne". Beregn sandsynligheden for at få 2 sorte, 3 hvide og 1 rød kugle.

LØSNING:

Lad X_1 være antallet af sorte kugler, X_2 være antallet af hvide kugler og X_3 være antallet af røde kugler.

Analogt med begrundelsen for den hypergeometriske fordeling fås:

$$P(X_1=2, X_2=3, X_3=1)=\frac{K(5,2)\cdot K(4,3)\cdot K(3,1)}{K(12,6)}=\frac{10\cdot 4\cdot 3}{924}=\underline{\underline{0.13}}$$

◆

8.7 Polynomialfordelingen.

Binomialfordelingen benyttes som model ved uafhængige gentagelser af samme eksperiment. Eksperimentet har to udfald succes eller ikke succes og der er en konstant sandsynlighed for succes. Hvis der foreligger flere end to udfald, f.eks. udtagning af møtrikker fra en løbende produktion, hvor diameter enten tilhører et givet toleranceinterval eller er for stor eller for lille, kan man generalisere til polynomialfordelingen. Idet formlen for binomialfordelingen kan skrives

$$f(x)=\binom{n}{x}\cdot p^x\cdot(1-p)^{n-x}=\frac{n!}{x!\cdot(n-x)!}\cdot p^x\cdot(1-p)^{n-x}=\frac{n!}{x_1!\cdot x_2!}\cdot p_1^{x_1}\cdot p_2^{x_2}\text{, hvor}$$

$p_1+p_2=1$ og $x_1+x_2=n$ fås analogt

DEFINITION af polynomialfordeling.

Lad n være et positivt helt tal, og lad $p_1+p_2+\ldots+p_k=1$ *og* $x_1+x_2+\ldots+x_k=n$ *hvor alle pér er positive tal og alle xér er hele tal.*

Sandsynlighedsfordelingen for en polynomialfordelt stokastisk variabel $(X_1, X_2, \ldots, X_k)$ *er*

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{n!}{x_1! \cdot x_2! \cdot \ldots \cdot x_k!} p_1^{x_1} \cdot p_2^{x_2} \cdot \ldots \cdot p_k^{x_k}$$

Dette illustreres ved følgende eksempel:

Eksempel 8.14. Polynomialfordelingen

En stor produktion af glaskugler indeholder 40% sorte, 35% hvide og 25% røde kugler.

Vi betragter det tilfældige eksperiment: "Udtrækning af 6 kugler observation af farven på kuglerne".

Beregn sandsynligheden for at få 2 sorte, 3 hvide og 1 rød kugle.

LØSNING:

Lad X_1 være antallet af sorte kugler, X_2 være antallet af hvide kugler og X_3 være antallet af røde kugler.

Vi får nu $$P(X_1 = 2, X_2 = 3, X_3 = 1) = \frac{6!}{2! \cdot 3! \cdot 1!} 0.4^2 \cdot 0.35^3 \cdot 0.25^1 = \underline{0.1029}$$

8.8. OVERSIGT over centrale formler i kapitel 8

X er **binomialfordelt** b(n,p), hvor n er kendt og p ukendt. Givet stikprøveværdi x

Konfidensinterval

Forudsætninger	Estimat for p	100 (1 - α) % konfidensinterval for parameter
eksakt	$\tilde{p} = \frac{x}{n}$	se eksempel 8.9
approksimation med normal-fordeling $10 \le x \wedge x \le n-10$	$\tilde{p} = \frac{x}{n}$	$\tilde{p} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}} \le p \le \tilde{p} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}$ se eksempel 8.9

Test af parameter p for binomialfordelt variabel

Der foreligger en stikprøve på X. Observeret stikprøveværdi x. Signifikansniveau er α.
Y er binomialfordelt $b(n, p_0)$, hvor p_0 er en given konstant

Alternativ hypotese H	*P* - værdi	Beregning	H_0 forkastes
$H: p > p_0$	$P(Y \ge x)$	1-BINOMIAL.FORDELING(x-1;n;p_0,1)	*P*-værdi $< \alpha$
$H: p < p_0$	$P(Y \le x)$	BINOMIAL.FORDELING(x;n;p_0,1)	
$H: p = p_0$	$P(Y \ge x)$ for $x > n \cdot p_0$ $P(Y \le x)$ for $x \le n \cdot p_0$	som række 1 som række 2	*P*-værdi $< \frac{1}{2}\alpha$

X er Poissonfordelt med middelværdi μ , hvor μ er ukendt.

Der optælles i alt *m* impulser i en stikprøve

Konfidensinterval

Forudsætning	Estimat for parameter	100 (1 - α) % konfidensinterval for parameter
Approksima-tion m ≥ 10	μ = m	$m - z_{1-\frac{\alpha}{2}}\sqrt{m} \le \mu \le m + z_{1-\frac{\alpha}{2}}\sqrt{m}$ se eksempel 8.12

Test af parameter μ for Poissonfordelt variabel .

Der optælles i alt *m* impulser i en stikprøve. Signifikansniveau er α.
Y er Poissonfordelt $p(\mu_0)$, hvor μ_0 er en given konstant.

Alternativ hypotese H	*P* - værdi	Beregning	H_0 forkastes
$H: \mu > \mu_0$	$P(Y \ge m)$	1-POISSON(m-1; μ_0 ;1)	*P* - værdi $< \alpha$
$H: \mu < \mu_0$	$P(Y \le m)$	POISSON(m; μ_0 ;1)	
$H: \mu \ne \mu_0$	$P(Y \ge m)$ for $x \ge \mu_0$ $P(Y \le m)$ for $x < \mu_0$	som række 1 som række 2	*P*-værdi $< \frac{1}{2}\alpha$

OPGAVER

Opgave 8.1

Ved en lodtrækning fordeles 3 gevinster blandt 25 lodsedler. En spiller har købt 5 lodsedler. 1) Beregn sandsynligheden for at spilleren vinder netop én gevinst.

Lad den stokastiske variable X være bestemt ved

X = antal gevinster som spilleren vinder

2) Find og skitser tæthedsfunktionen for X

3) Beregn middelværdien for X

Opgave 8.2

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 8.3

På en undervisningsinstitution skal 105 studerende holde fest sammen med deres 23 lærere. Et festudvalg på 3 personer vælges tilfældigt. Beregn sandsynligheden for at der kommer både lærere og studerende med i udvalget.

Opgave 8.4.

I en kortbunke er der 26 kort, hvoraf netop 4 er spar. Kortene fordeles i 2 lige store bunker A og B.

1) Peter påstår, at sandsynligheden for at bunke A indeholder netop 3 spar er 24.87%. Har Peter ret?

2) Beregn sandsynligheden for, at en af bunkerne indeholder netop 1 spar.

Opgave 8.5

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.

Opgave 8.6

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 8.7
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 8.8
En "sypigetipper" (M/K) deltog i tipning 42 gange i løbet af et år. På hver tipskupon var der 13 kampe, ved hver af hvilke tipperen ved systematisk gætning satte et af de 3 tegn: 1, x, 2. Beregn sandsynligheden p for, at tipperen det pågældende år tippede mindst 200 kampe rigtigt.

Opgave 8.9
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 8.10.
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; vi vil senere formulere dette således, at produktionen er **i statistisk kontrol** med fejlsandsynligheden $p = 2\%$.
Hvor stor er sandsynligheden for, at partiet under disse omstændigheder accepteres?

Opgave 8.11
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 8.12
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 8.13
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 8.14
En maskinfabrikant påtænker at købe 100000 møtrikker af en bestemt type. Man beslutter sig til at købe et tilbudt parti af den nævnte størrelse, såfremt en stikprøve på 150 møtrikker højst indeholder 4% defekte møtrikker.

1) Beregn sandsynligheden for, at partiet bliver godkendt af maskinfabrikken, såfremt det indeholder
   a) 4% defekte møtrikker,
   b) 7,5% defekte møtrikker,
2) Bestem, for hvilken procentdel defekte møtrikker det ovennævnte parti har 50% sandsynlighed for at blive godkendt af maskinfabrikken.

Opgave 8.15
En ny vaccine formodes med en sandsynlighed på mindst 85% at have en forebyggende virkning over for en bestemt influenzatype.
Før en truende influenzaepedemi vaccineres et hospitalspersonale på 600 personer med den pågældende vaccine. 125 af disse bliver smittet af sygdommen.
Kan dette opfattes som en eksperimentel påvisning af, at vaccinen er mindre virksom end ventet?

Opgave 8.16

1) Antag, at en vis type af fostermisdannelse normalt forekommer med hyppigheden 164 tilfælde p. 100000 fødsler. Beregn sandsynligheden for 3 eller flere fostermisdannelser blandt 256 fødsler.
2) For at undersøge om forholdene i et bestemt arbejdsmiljø forøger hyppigheden af denne type misdannelse, undersøgte man hyppigheden af misdannelser for mødre, som under graviditeten havde haft den aktuelle type af arbejde, og fandt 3 misdannelser blandt 256 fødsler. Kan den forøgede relative hyppighed i dette materiale skyldes tilfældigheder?

Opgave 8.17
Udsættes planterne af en bestemt sort roser for meldugssmitte, bliver i middel brøkdelen p angrebet, hvor p er mindst 0.20. En rosengartner fremavler en rosenstamme, som han påstår er mere modstandsdygtig over for meldugssmitte. For at kontrollere denne påstand bliver 100 roser af den nye stamme udsat for meldugssmitte. Det viser sig, at 12 roser bliver angrebet.

1) Bekræfter dette resultat rosengartnerens påstand? (Husk altid at anføre: Hvad X er. Antagelser. Nulhypotese. Beregninger. Konklusion.).

Hvis rosengartneren har ret, skal man

2) Angiv et estimat $\tilde{p}$ for den nye stammes p.
3) Angiv et 95% konfidensinterval for den nye stammes p.

Opgave 8.18
En fabrikant af chip til computere reklamerer med, at højst 2% af en bestemt type chip, som fabrikken sender ud på markedet er defekte.
Et stort computerfirma vil købe et meget stort parti af disse chip, hvis påstanden er rigtigt. For at teste påstanden købes 1000 af dem. Det viser sig, at 33 ud af de 1000 er defekte.
Kan fabrikantens påstand på denne baggrund forkastes på signifikansniveau 5% ?

Opgave 8.19

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.

Opgave 8.20

Af 1000 tilfældigt udvalgte patienter, der led af lungekræft, var 823 døde senest 5 år efter sygdommen blev opdaget.

Angiv på dette grundlag et 95% konfidensinterval for sandsynligheden for at dø af denne sygdom senest 5 år efter at sygdommen bliver opdaget.

Opgave 8.21

En fabrikant af lommeregnere vurderer, at ca. 1% af de producerede lommeregnere er defekte. For at få en nøjere vurdering heraf ønskes udtaget en stikprøve, der er så stor, at radius i et 95% konfidensinterval for fejlprocenten p er højst 0.5%.

Find stikprøvens størrelse n.

Opgave 8.22

På en fabrik fremstilles gulvtæpper, som har størrelsen 20 m^2. Ved fabrikationen er der gennemsnitlig 6 vævefejl pr. 100 m^2 klæde.

1) Beregn sandsynligheden for, at et tilfældigt gulvtæppe ingen vævefejl har.
2) Beregn sandsynligheden for, at et tilfældigt gulvtæppe højst har 2 vævefejl.

Fabrikken køber en ny væv. For at få et estimat for middelværdien måltes antallet af vævefejl i 12 gulvtæpper hver på 20 m^2. Resultaterne var

Gulvtæppe nr	1	2	3	4	5	6	7	8	9	10	11	12
Antal vævefejl	4	2	7	3	4	5	5	8	1	1	3	5

3) Find et estimat for middelværdien af antal vævefejl p. 20 m^2 klæde.

Opgave 8.23

Et radioaktivt præparat undergår gennemsnitligt 100 desintegrationer (sønderdelinger) p. minut. Lad X betegne antal desintegrationer i et sekund (som er lille i forhold til præparatets halveringstid).

Find $P(X \leq 1)$.

Opgave 8.24

Ved en TV-fabrikation optælles som led i en godkendelseskontrol antal loddefejl p. 5 TV-apparater. Fabrikanten ønsker at få et overblik over antal loddefejl, og optalte derfor antal loddefejl på 24 tilfældigt udtagne TV apparater. Resultatet fremgår af skemaet:

Antal loddefejl	0	1	2	3	4	5	6	7	8	9
Antal TV apparater	3	2	4	6	5	2	1	0	1	0

Lad X være antallet af loddefejl i 5 TV apparater.

1) Angiv den sandsynlighedsfordeling X approksimativt kan antages at følge, og giv et estimat for parameteren i fordelingen.
2) Beregn på basis af svaret i spørgsmål 1 sandsynligheden for, at der på 5 tilfældigt udtagne TV-apparater højst er i alt 18 loddefejl?

Opgave 8.25

På et teknisk universitet er et centralt edb-anlæg i konstant brug. Man har erfaring for, at anlægget i løbet af en 20 ugers periode har gennemsnitligt 7 maskinstop.
Beregn sandsynligheden p for, at anlægget i en 4 ugers periode har mindst ét maskinstop.

Opgave 8.26

På en fabrik indtræffer i gennemsnit 72 ulykker om året. Antag, at de forskellige ulykker indtræffer uafhængigt af hinanden, og at de er nogenlunde jævnt fordelt over året.
Beregn, idet et arbejdsår sættes lig med 48 uger, sandsynligheden for at der i en uge indtræffer flere end 3 ulykker.

Opgave 8.27

Til et bestemt telefonnummer er der i løbet af aftenen i middel 300 opkald i timen.
Beregn sandsynligheden for, at der i løbet af et minut er højst 8 opkald.

Opgave 8.28

En fabrikation af fortinnede plader finder sted ved en kontinuerlig elektrolytisk proces. Umiddelbart efter produktionen kontrolleres for pladefejl. Man har erfaring for, at der i middel er 1 pladefejl hvert 5'te minut.
Beregn sandsynligheden for, at der højst er 5 pladefejl ved en halv times produktion.

Opgave 8.29

Lastbiler med affald ankommer tilfældigt og indbyrdes uafhængigt til en losseplads. Lossepladsens maksimale kapacitet er beregnet til, at der i middel ankommer 90 lastbiler p. time. Ledelsen af pladsen føler, at travlheden er blevet større i den sidste tid, således at antallet af lastbiler overskrider den maksimale kapacitet. For at undersøge dette, foretages en optælling af lastbiler i perioder à 10 minutter. Følgende resultater fremkom:

14	17	18	16	18	12	22	16	21	18

1) Bekræfter disse resultater ledelsens formodning? (Husk altid at anføre: Hvad X er. Antagelser. Nulhypotese. Beregninger. Konklusion.).

Forudsat man kan vise, at ledelsen har ret på et signifikansniveau på 5%, skal man

2) Angiv et estimat $\tilde{\mu}$ for middelværdien μ [lastbiler/time].
3) Angiv et 95% konfidensinterval for middelværdien μ [lastbiler/time].

Opgave 8.30

Nedenstående tabel viser fordelingen af 400 volumenenheder med hensyn til antal gærceller p. volumenenhed.

Antal gærceller	0	1	2	3	4	5	6	7	8	9	10	11	12
Antal volumenenheder	0	20	43	53	86	70	54	37	18	10	5	2	2

Lad X være antal gærceller p. volumenenhed. Det antages, at X er en stokastisk variabel der er Poissonfordelt p (μ).

1) Find et estimat $\tilde{\mu}$ for μ .
2) Angiv et 95% konfidensinterval for μ .
3) Forudsat at X er Poissonfordelt p ($\tilde{\mu}$) ønskes beregnet det forventede antal volumenenheder, hvori der forekommer 5 gærceller (for $x = 5$).

Opgave 8.31

Ved inspektion af en produktion med isolering af kobberledning taltes der i løbet af 50 minutter i alt 11 isoleringsfejl.

Idet antallet af isoleringsfejl p. 50 minutter antages at være Poissonfordelt p (μ_1), skal man

1a) angive et estimat for μ_1 .
1b) angive et 95% konfidensinterval for μ_1 .

Det oplyses nu, at man i hver 5 minutters periode i den ovenfor omtalte 50 minutters periode havde observeret følgende antal isoleringsfejl:

Periode	1	2	3	4	5	6	7	8	9	10
Antal fejl	1	0	2	2	1	1	3	0	1	0

Idet antallet af isoleringsfejl p. 5 minutter antages at være Poissonfordelt p (μ_2), skal man

2a) angive et estimat for μ_2 .
2b) angive et 95% konfidensinterval for μ_2 .

Opgave 8.32

I en urne findes 10 røde kugler, 5 hvide kugler og 3 sorte kugler.6 gange efter hinanden optages tilfældigt en kugle fra urnen. Bestem sandsynligheden for, at der i alt er optaget 1 rød, 2 hvide og 3 sorte kugler, når

1) kuglerne optages uden tilbagelægning
2) kuglerne optages med tilbagelægning.

Opgave 8.33

En virksomhed fabrikerer farvede glasklodser til dekorationsbrug. Defekte glasklodser frasorteres. Man har erfaring for, at af de frasorterede klodser har i middel 50% kun revner, 35% kun farvefejl, medens resten har begge disse fejl.
Beregn sandsynligheden for, at af 12 tilfældige defekte klodser har 6 kun revner, 4 kun farvefejl og 2 begge disse fejl.

Opgave 8.34

I et kortspil med de sædvanlige 52 spillekort har en spiller modtaget 13 kort. Angiv i procent med 2 decimaler sandsynligheden for, at 3 af disse er esser og 5 er billedkort.

9 ANDRE KONTINUERTE FORDELINGER

9.1 INDLEDNING

Vi vil i dette kapitel kort orientere om en række fordelinger, som er vigtige i specielle sammenhænge,

9.2 DEN REKTANGULÆRE FORDELING

DEFINITION af rektangulær fordeling med parametrene a og b.

Lad a og b være to reelle tal, hvor a<b.

Sandsynlighedsfordelingen for en kontinuert stokastisk variabel X med tæthedsfunktionen f (x)

bestemt ved $$f(x)=\begin{cases}\dfrac{1}{b-a} & \text{for} \quad a\le x\le b\\ 0 & \text{ellers}\end{cases}$$

siges at være rektangulært fordelt rekt (*a , b*).

SÆTNING 9.1. (*Middelværdi og spredning for rektangulær fordeling*).

Den rektangulære fordeling har $E(X)=\dfrac{a+b}{2}$ *og* $\sigma(X)=\dfrac{b-a}{2\cdot\sqrt{3}}$ *(a < b)*

Bevis:

$$E(X)=\int_a^b \frac{x}{b-a}dx=\left[\frac{x^2}{2\cdot(b-a)}\right]_a^b=\frac{b+a}{2}$$

$$V(X)=\int_a^b \frac{\left(x-\frac{a+b}{2}\right)^2}{b-a}dx=\left[\frac{\left(x-\frac{a+b}{2}\right)^3}{3(b-a)}\right]_a^b=\frac{(b-a)^2}{12}$$

◆

Eksempel 9.1 Kontinuert variabel.

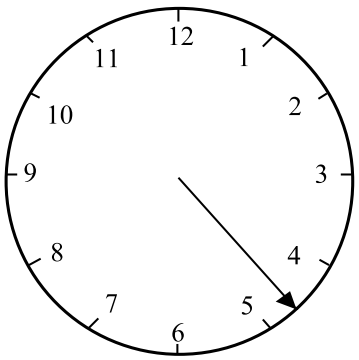
Lad randen af en roulette være ækvidistant inddelt efter en skala fra 0 til 12, jævnfør figuren.

Ved et roulettespil bringes roulettens viser til at rotere, hvorefter den standser ud for et tilfældigt punkt på skalaen.

Lad *X* være det tal som roulettens viser peger på.

Idet *X* må kunne antage ethvert tal mellem 0 og 12, må *X* være en kontinuert variabel.

Angiv tæthedsfunktion og fordelingsfunktion for *X* og skitser disse.

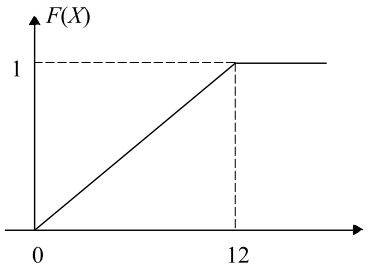


Løsning:

Da $P(0 \le X \le x) = \frac{x}{12}$ for $0 \le x \le 12$

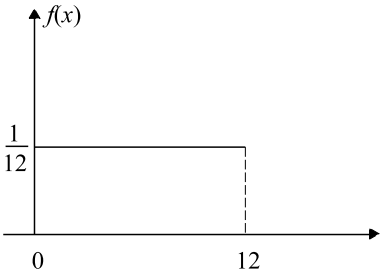
er fordelingsfunktionen for X

$$F(x) = \begin{cases} 0 & \text{for} \quad x \le 0 \\ \frac{x}{12} & \text{for} \quad 0 \le x \le 12 \\ 1 & \text{for} \quad x \ge 12 \end{cases}$$



Ved differentiation fås tæthedsfunktionen

$$f(x) = \begin{cases} \frac{1}{12} & \text{for} \quad 0 \le x \le 12 \\ 0 & \text{ellers} \end{cases}$$



9.3 EKSPONENTIALFORDELINGEN

I kapitel 7 betragtede vi antallet N af revner pr. meter langs et kobberkabel. Vi antog, at N var Poissonfordelt. Hvis vi i stedet havde betragtet afstanden X mellem revnerne, havde vi fået en ny stokastisk variabel, som må være kontinuert. Som det fremgår af følgende sætning er X eksponentialfordelt.

SÆTNING 9.2 .*Eksponentialfordeling*.

Lad W være en Poissonfordelt stokastisk variabel.

Lad det gennemsnitlige antal impulser i en tidsenhed være λ. Lad X være tiden indtil næste impuls.

X er da en kontinuert stokastisk variabel med sandsynlighedsfordelingen (tæthedsfunktionen) $f(x) = P(X = x)$ bestemt ved

$$f(x) = \begin{cases} \frac{1}{\mu} \cdot e^{-\frac{x}{\mu}} & \text{for} \quad x > 0 \\ 0 & \text{ellers} \end{cases} \qquad \text{hvor} \quad \mu = \frac{1}{\lambda}$$

X siges at være eksponentialfordelt exp (μ) med parameteren μ.

Middelværdien for exp (μ) er $E(X) = \mu$ og spredningen er $\sigma(X) = \mu$.

Bevis:

I tidsrummet fra x_0 til $x_0 + x$ er der I gennemsnit $\lambda \cdot x$ impulser. Lad W være det aktuelle antal impulser i tidsrummet $[x_0 ; x_0 + x]$. W er da Poissonfordelt $p(\lambda \cdot x)$.

Idet X er tiden fra én impuls til den næste, er $P(X > x) = P(W = 0)$, da der ingen impulser er i tidsrummet $[x_0 ; x_0 + x]$.

Da $P(W=0)=\frac{(\lambda\cdot x)^0}{0!}\cdot e^{-\lambda\cdot x}=e^{-\lambda\cdot x}$, er $P(X>x)=e^{-\lambda\cdot x}$.

Vi har derfor $F(x)=P(X\le x)=1-P(X>x)=1-e^{-\lambda\cdot x}$.

Ved differentiation fås tæthedsfunktionen: $f(x)=F'(x)=\lambda\cdot e^{-\lambda\cdot x}$. Sættes $\lambda=\frac{1}{\mu}$ fås formlen. ◆

Bevis for middelværdi og spredning:

$$E(X)=\int_0^\infty \lambda\cdot x\cdot e^{-\lambda\cdot x}dx=\left[-e^{-\lambda\cdot x}\left(x-\frac{1}{-\lambda}\right)\right]_0^\infty=\frac{1}{\lambda}=\mu$$

$$V(X)=E(X^2)-(E(X))^2=\int_0^\infty \lambda\cdot x^2\cdot e^{-\lambda\cdot x}dx-\mu^2=\left[-e^{-\lambda\cdot x}\left(x^2-\frac{2x}{\lambda}+\frac{2}{\lambda^2}\right)\right]_0^\infty-\mu^2=\frac{2}{\lambda^2}-\mu^2=\mu^2.$$ ◆

Som det fremgår af beviset for sætning 9.2, er fordelingsfunktionen for en eksponentialfordelt variabel bestemt ved udtrykket

$$F(x)=P(X\le x)=\begin{cases}1-e^{-\frac{x}{\mu}} & \text{for} \quad x>0\\ 0 & \text{ellers}\end{cases}$$

På nedenstående graf er afbildet tæthedsfunktionen for eksponentialfordelingerne exp(1.0) og exp(2.0)

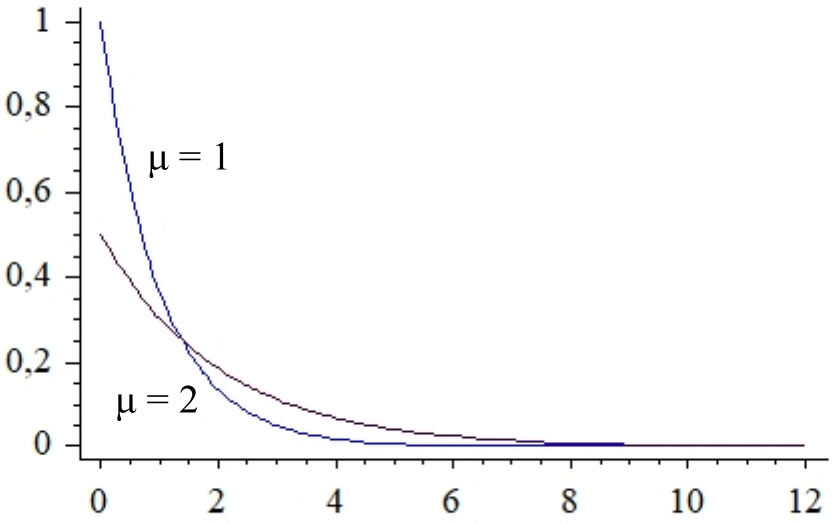


Fig 9.1 *Eksponentialfordelingerne* exp(1) *og* exp(2)

Eksempel 9.2. Afstanden mellem successive revner i kabel.

Vi betragter det i eksempel 9.10 omtalte problem, hvor man fandt, at antallet N af mikroskopiske revner i et kobberkabel er Poissonfordelt. Der var i gennemsnit 12.3 af den type revner pr. 10 meter. Lad X være afstanden mellem to på hinanden følgende revner.

Beregn sandsynligheden for, at der er mere end 1 meter mellem to revner.

Løsning

Da der i gennemsnit er 1.23 revner pr. meter, må der i gennemsnit være $\frac{1}{1.23} = 0.812$ meter mellem to revner. Vi har derfor at X er eksponentialfordelt med $\mu = 0.813$.

$$P(X > 1) = 1 - P(X \le 1) = 1 - \left(1 - e^{-\frac{1}{0.813}}\right) = e^{-1.23} = \underline{\underline{0.2923}}$$ ◆

Levetider. I apparater, som består af elektroniske komponenter (eksempelvis lommeregnere), er der et meget ringe mekanisk slid. Apparatets fremtidige levetid vil derfor (næsten ikke) afhænge af, hvor længe det har fungeret indtil nu. I sådanne tilfælde vil eksponentialfordelingen erfaringsmæssigt være en god approksimativ model for apparatets levetid.

Det kan nemlig vises, at eksponentialfordelingen er den eneste kontinuerte fordeling, som har ovennævnte egenskab (er uden hukommelse)

Bevis: Lad X være eksponentialfordelt med middelværdi μ og lad $b > a > 0$ være vilkårlige konstanter. Der gælder da:

$$P(X > a+b \mid X > a) = \frac{P\big((X > a+b) \wedge (X > a)\big)}{P(X > a)} = \frac{P(X > a+b)}{P(X > a)} = \frac{e^{-\frac{a+b}{\mu}}}{e^{-\frac{b}{\mu}}} = e^{-\frac{b}{\mu}} = P(X > b)$$ ◆

Eksempel 9.3. Levetid for elektriske pærer.

Man har erfaring for, at en bestemt type elektriske pærer har en "brændtid" T (målt i timer), som approksimativt er eksponentialfordelt. På basis af et stort antal målinger ved man , at middellevetiden er $\mu = 1500$ timer.

1) Hvor stor er sandsynligheden for, at en tilfældig pære brænder over, inden den har været tændt i 1200 timer?
2) Find sandsynligheden for, at en tilfældig pære brænder i mere end 1800 timer.
3) En pære har brændt i 800 timer. Hvad er sandsynligheden for, at den brænder i mindst 1800 timer mere.

Løsning

1) $P(T<1200)=F(1200)=1-e^{-\frac{1200}{1500}}=1-0.449=\underline{\underline{55.1\%}}$.

2) $P(T>1800)=1-F(1800)=e^{-\frac{1800}{1500}}=\underline{\underline{30.12\%}}$

3) Da eksponentialfordelingen ingen hukommelse har, vil svaret blive som i spørgsmål 2, dvs. $\underline{30.12\%}$.

◆

9.4 WEIBULLFORDELINGEN

Hvis komponenterne i et elektronisk apparat ikke "slides", dvs. den fremtidige levetid ikke afhænger af den foregående tid, er som nævnt i afsnit 9.3 eksponentialfordelingen velegnet som model for apparatets levetid.

Hvis derimod de pågældende komponenters eventuelle svigten afhænger af den forløbne tid, kan man ofte med fordel benytte den i det følgende nævnte **Weibullfordeling** som approksimativ model for apparatets levetid (model for apparatets pålidelighed).

DEFINITION af Weibulfordeling. *Lad k og μ være positive tal. Sandsynlighedsfordelingen for en kontinuert stokastisk variabel X med tæthedsfunktionen f (x) bestemt ved*

$$f(x)=\begin{cases}\dfrac{k}{\mu^k}\cdot x^{k-1}\cdot e^{-\left(\frac{x}{\mu}\right)^k} & for \quad x>0\\ 0 & ellers\end{cases}$$

siges at være Weibullfordelingen $wei(k,\mu)$.

Det kan vises, at Weibullfordelingen $wei(k,\mu)$ har middelværdien $E(X)=\mu\cdot\Gamma\left(\frac{k+1}{k}\right)$ [1)]

og spredningen $\sigma(X)=\mu\cdot\sqrt{\Gamma\left(\frac{k+2}{k}\right)-\left(\Gamma\left(\frac{k+1}{k}\right)\right)^2}$

Det ses, at Weibullfordelingen kan opfattes som en generalisation af eksponentialfordelingen, idet $wei(1,\mu)=\exp(\mu)$.

Såfremt levetiderne for komponenter i et apparat aftager jo længere tid apparatet har været i funktion (på grund af slid), kan man benytte en Weibullfordeling med $k>1$ som approksimativ model for apparatets levetid.

---

1) Gammafunktionen $\Gamma(x)$ er defineret i "Supplement til statistiske grundbegreber" 3A

9.5 DEN LOGARITMISKE NORMALFORDELING

Indenfor det biokemiske eller biologiske område (forsøgsdyrs reaktionstid, cellevækst m.v.) er den stokastiske variabel X ikke normalfordelt, men hvis man foretager en logaritmisk transformation $Y = \ln X$ er Y (approksimativt) normalfordelt.

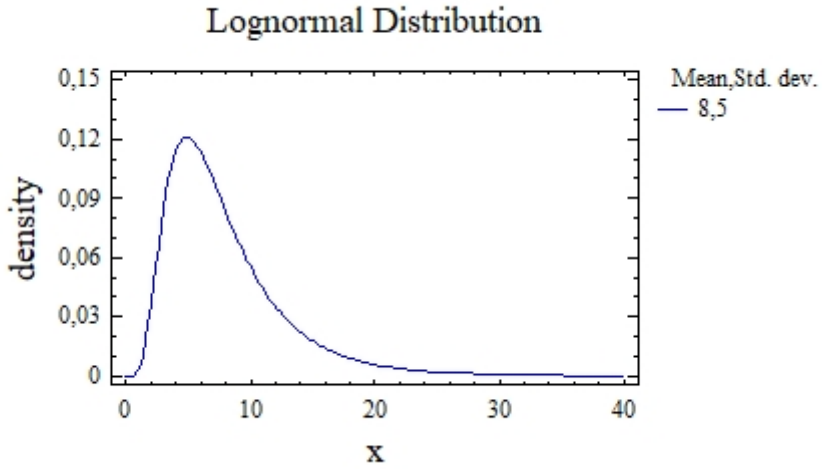
Man siger så, at X er **logaritmisk normalfordelt.**

Tæthedsfunktionen for X er bestemt ved $f(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma}\frac{1}{x}e^{-\frac{1}{2}\left(\frac{\ln x-\mu}{\sigma}\right)^2}$ for $x > 0$.

Det kan vises, at mens $Y = \ln X$ har middelværdi μ og spredning σ har X middelværdi

$E(X) = e^{\mu}\cdot e^{-\frac{1}{2}\cdot\sigma^2}$ og $V(X) = e^{2\mu}\cdot e^{\sigma^2}\cdot(e^{\sigma^2}-1)$.

Nedenfor er tegnet en logaritmisk normalfordeling med middelværdi 8 og spredning 5.

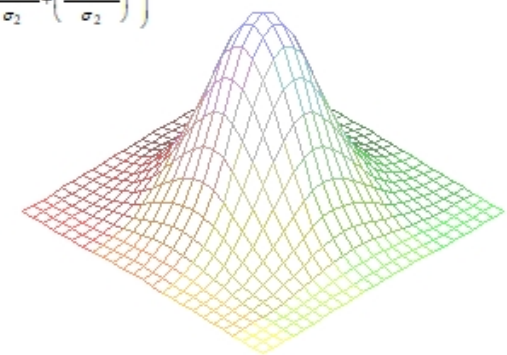


9.6 DEN 2-DIMENSIONALE NORMALFORDELING

Flerdimensionale fordelinger vil blive omtalt nærmere i kapitel 12. Her nævnes uden forklaring et eksempel herpå.

DEFINITION af 2-dimensional normalfordeling Lad μ_1, μ_2 være reelle tal og σ_1, σ_2 *være positive tal. Sandsynlighedsfordelingen for 2-dimensional kontinuert stokastisk variabel* (X_1, X_2) *med tæthedsfunktion bestemt ved*

$$f(x) = \frac{1}{2\pi\cdot\sigma_1\cdot\sigma_2\sqrt{1-\rho^2}}\cdot e^{-\frac{1}{2\cdot(1-\rho^2)}\left(\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2-2\rho\frac{x_1-\mu_1}{\sigma_1}\cdot\frac{x_2-\mu_2}{\sigma_2}+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right)}$$



kaldes den 2-dimensionale ***normalfordeling*** *med parametrene* μ_1, μ_2, σ_1 og σ_2.

Det kan vises, at $E(X_1) = \mu_1$, $E(X_2) = \mu_2$, $\sigma(X_1) = \sigma_1, \sigma(X_2) = \sigma_2$ og $\rho(X_1, X_2) = \rho$

Grafen ses overfor.

OPGAVER

Opgave 9.1

På et betalingsnummer måltes man i tidsrummet fra kl 20 til 22 tiden t (antal minutter) mellem på hinanden følgende telefonopkald. Følgende resultater fandtes:

Beliggenhed af t	]0;1]	]1;2]	]2;3]	]3;4]	]4;5]	]5;6]	]6;7]	]7;8]	]8;9]	]9;10]	]10; ∞ [
Antal observationer	36	21	16	13	7	9	6	1	2	6	0

Det antages, at antallet N af telefonopkald til nummeret er Poissonfordelt. Lad T være tiden mellem to opkald.

1) Angiv fordelingsfunktionen for T, og giv et estimat for middelværdien μ.
   Vink: Antage, at for alle observationer i et interval er tidsrummet mellem observationerne intervallets midterværdi.
2) På baggrund af den i spørgsmål 1 fundne estimat for μ, ønskes bestemt $P(2 < T \leq 3)$.
3) Af tabellen ses, at i intervallet]2; 3] forekommer i alt 16 observationer. Angiv hvor mange observationer man må forvente, ud fra resultatet i spørgsmål 2.

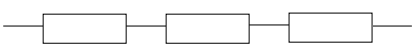
Opgave 9.2

Om en bestemt type elektriske komponenter vides, at deres levetider er eksponentialfordelte med en middellevetid på 800 timer.

1) Find sandsynligheden for, at en komponent holder mindst 200 timer.
2) Find sandsynligheden for, at en komponent holder mellem 600 og 800 timer.
3) En komponent har holdt i 900 timer. Find sandsynligheden for, at den kan holde i mindst 200 timer mere.
4) I et elektrisk system indgår netop én komponent af denne type. Hver gang komponenten svigter, udskiftes den øjeblikkeligt med en ny komponent af samme type. Find sandsynligheden for, at komponenten udskiftes 12 gange i løbet af 8000 timer.

Opgave 9.3

Antag, at levetiderne for en bestemt slags elektroniske komponenter er uafhængige og alle er eksponentialfordelt



med en middellevetid på 3 (år). Betragt et delsystem bestående af 3 sådanne komponenter i seriekobling:(en seriekobling ophører at fungere, når én af komponenterne ophører at fungere). Bestem middellevetiden for et sådant system.

Opgave 9.4

Nedbrydningstiden i den menneskelige organisme for et givet kvantum af et bestemt stof antages at være eksponentialfordelt med middelværdien 5 timer.

Ved et forsøg indsprøjtes stoffet samtidig i 10 patienter.

1) Beregn sandsynligheden (afrundet til et helt antal procent) for, at stoffet hos en tilfældig valgt patient vil være nedbrudt efter 8 timers forløb.
2) Beregn sandsynligheden for, at stoffet efter 8 timers forløb vil være nedbrudt hos mindst 5 af patienterne.
3) Efter hvor mange timers forløb vil der være ca. 90% sandsynlighed for, at stoffet er nedbrudt hos samtlige 10 patienter?
4) Hvor mange patienter skal indgå i en ny undersøgelse, hvis der skal være ca. 95% sandsynlighed for, at der er mindst en patient, hvis organisme efter 8 timers forløb endnu ikke har nedbrudt stoffet?

10. GRUNDLÆGGENDE OPERATIONER I Excel

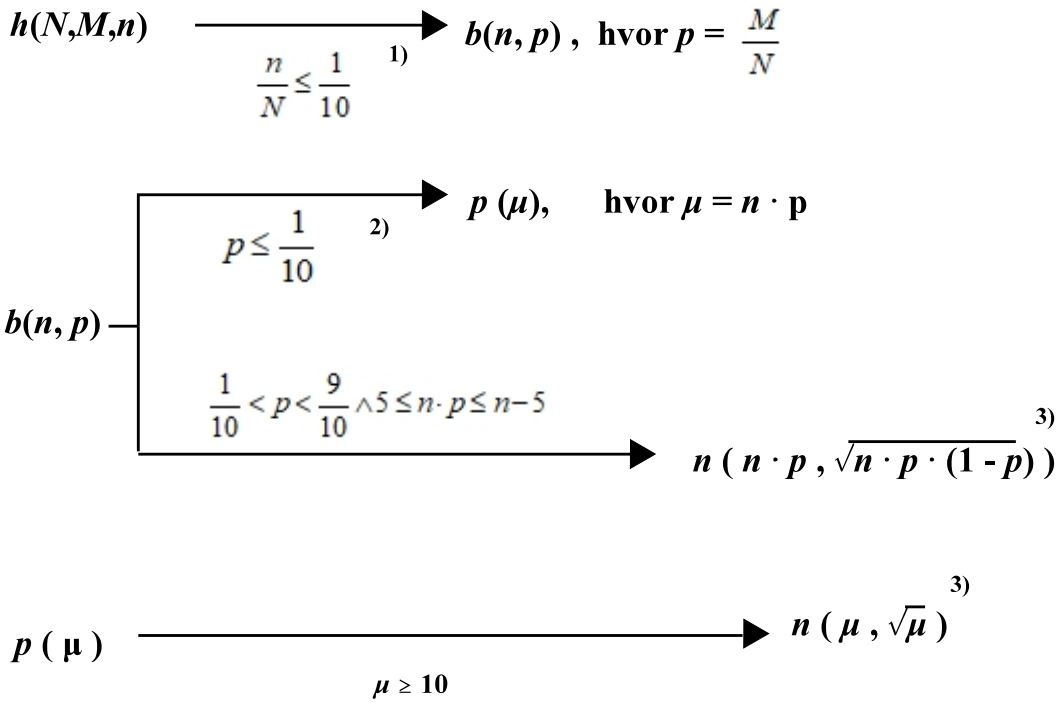
Forudsætninger.

Da ikke alle de anvendte statistiske funktioner er indbygget fra starten, skal man først vælge et tilføjelsesprogram:
Vælg "Filer" ▶ Indstillinger ▶ Tilføjelsesprogrammer ▶ marker Analysis toolpak ▶ Udfør ▶ Analysis toolpak VBA ▶ Udfør ▶ Problemløser ▶ Udfør.

Indlægge rækkenumre og søjlenumre

Når man skal forklare ordrer i Excel kan det være godt, at række- og søjlenumre indgår i udskrift:
Vælg sidelayout ▶ under overskrifter og gitterlinier: marker vis og udskrift.

APPENDIX . Oversigt over approksimationer.



1) Når $\frac{n}{N} > \frac{1}{10}$ og $\frac{M}{N} \leq \frac{1}{10}$ benyttes, at $h(N, M, n) = h(N, n, M)$.

2) For $p \geq \frac{9}{10}$ benyttes, i stedet for at tælle X_{gammel} = "antal af successer", så at tælle X = antal fiaskoer dvs. $p = 1 - p_{\text{gammel}}$ og $X = n - X_{\text{gammel}}$.

3) Der skal foretages "heltalskorrektion".

Eksempel : Lad X være binomialfordelt eller Poissonfordelt:

a) Skal udregnes P(X ≤ 10). Ved approksimation med en normalfordeling, beregnes i normalfordeling P(X ≤10.5).

b) Skal udregnes P(X ≥ 10). Ved approksimation med en normalfordeling, beregnes i normalfordeling P(X ≥9.5).

FACITLISTE

KAPITEL 2
2.1 -
2.2 -
2.3
2.4 (1) - (2) ca 24%
2.5 (1) - (2) ca 0.05
2.6 (1) - (2) ca 13%
2.7 (1) - (2) 24.8 24.5
2.8 (1) - (2) - (3) -

KAPITEL 3
3.1 a) 65 0.4 b) 0.004

KAPITEL 4
4.1 (1) 0.7734 0.0548 0.1718 (2) 0.7480
4.2 (1) 69.15% (2) 10.88% (3) 112.2 (4) 117.3 6.535
4.3 (1) 86.64% (2) 0.00765 (3) [2.48 ; 2.52]
4.4 (1) 5.91% (2) 27.65% (3) [783.51; 816.49]
4.5 (1) 9.5 1.265 (2) 12.44 (3) 2.41%
4.6 (1) 92.75%
4.7 (1) 97.73% (2) 25.45
4.8 (1) 65 0.4 (2) 77.34%

KAPITEL 5
5.1 (1) 12.13 0.6783 (2) [11.64 ; 12.62] (3) [10.52 ; 13.74]
5.2 (1) 2259.92 35.569 (2) [2237.3 ; 2282.5] (3) [2178.4 ; 2341.4] (4) 34
5.3 (1) 74.0362 0.00124 (2) [74.035; 74.037]
5.4 (1) 19.13 (2) [14.0 ; 30.2] (3) 750.2 (4) [739.6 ; 760.8]
5.5 (1) 7.83 0.363 (2) [7.45 ; 8.22] (3) 51
5.6 (a) 4.256 [4.23 ; 4.29] (b) 66 (c) [4.16 ; 4.35] (d) [0.0285 ; 0.0756]
5.7 (1) [96.36 ; 111.14] (2) [2.638; 17.322]

KAPITEL 6
6.1 (1) nej P-værdi = 2.27%
6.2 (1) ja P-værdi = 3.6% (2) 58 [55.82; 60.18]
6.3 (1) ja P-værdi = 0.44% (2) 84.47 [81.26 ; 87.67] (4) 25.7%
6.4 (a) nej P-værdi = 12.1% (b) -
6.5 (a) ja P-værdi = 0.157% (b) [24.07 ; 35.56]
6.6 nej P-værdi = 6.44%
6.7 ja, P-værdi = $1.7 \cdot 10^{-9}$
6.8 (1) 35 (2) ja P-værdi = 0.16% (3) nej [2.59 ; 2.67]
6.9 (1) 35 (2) ja P-værdi = $1.9 \cdot 10^{-14}$ (3) ja

KAPITEL 7

7.1 0.5 0.8 0.2 0.7
7.2 (1) 0.9134 (2) 0.9678
7.3 (1) 8.75% (2) 38.75% (3) 41.25% (4)11.25%
7.4 (1) 6.4% (2) 78.4% (3) 7.2%
7.5 $1.283 \cdot 10^{12}$
7.6 (a) - (b) 736
7.7 (a) 6 (b) 24
7.8 (a) 100 /b) 2400
7.9 60
7.10 (a) 30.24% (b) 0.24% (c) 99.76% (d) 4.04% (e) 44.04% (f) 21.44%
7.11 (1) 27.1% 36.0% 9.756% (2) 53.34% (3) 49.20%
7.12 3^{40}
7.13 (1) 31 (2) 9
7.14 30.24%
7.15 $9 \cdot 10^{7}$

KAPITEL 8

8.1 (1) 41.3% 2)- 3) 0.6
8.2 (A) 0.018% (B) 1.29% (C) 38.24%
8.3 44.57%
8.4 (1) ja (2) 49.74%
8.5 (1) 17.68% (2) 59.28%
8.6 2.58%
8.7 5.83%
8.8 5.69%
8.9 95.47%
8.10 40.33%
8.11 12.85%
8.12 77.85%
8.13 (1) 7.93% (2) 11.8%
8.14 (1) (a) 60.63% (b) 6.17% (2) 4.44%
8.15 nej, P-værdi =0.0079%
8.16 (1) 0.9% (2) nej
8.17 (1) ja, P-værdi = 2.53% (2) 0.12 (3) [0.064 ; 0.200] (approx:[0.056 ; 0.184])
8.18 ja p = 0.43%
8.19 (1) 0.108 (2) [0.089 ; 0.129]
8.20 [0.798 ; 0.847]
8.21 1522
8.22 (1) 30.1% (2) 87.9% (3) 4
8.23 50.37%
8.24 (1) 15 (2) 81.9%
8.25 75.3%
8.26 6.56%

8.27 93.2%
8.28 44.57%
8.29 (1) ja P-værdi = 4.19% (2) 103.3 (4) [88.2 ; 118.8] (approx[87.78 ; 118.62])
8.30 (1) 4.68 (2) [4.47 ; 4.89] (3) 69.44
8.31 (1a) 11 (1b) [5 ; 18] eller [4.5 ; 17.5] (2a) 1.1 (2b) [0.5 ; 1.8] eller [0.45 ; 1.75]
8.32 (1) 0.539% (2) 0.119%
8.33 7.31%
8.34 0.188%

KAPITEL 9
9.1 (1) 2.90 (2) 14.6% (3) 16.98
9.2 (1) 77.88% (2) 10.45% (3) 77.88% (4) 9.48%
9.3 1
9.4 (1) 79.8% (2) 99.33% (3) 22.8 (4) 14

STIKORDSREGISTER