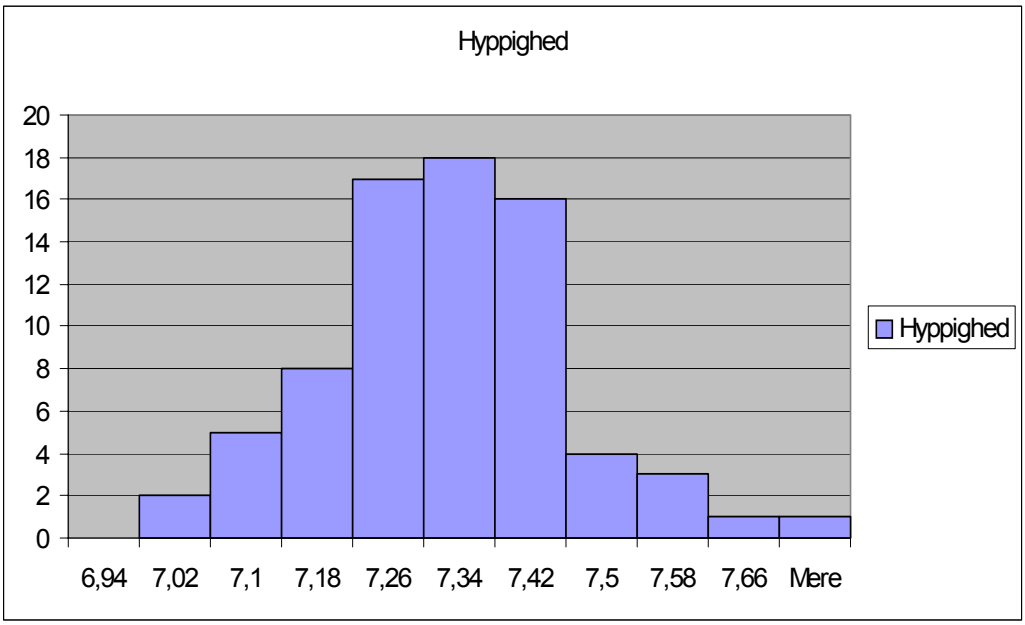
MOGENS ODDERSHEDE LARSEN

Grundlæggende STATISTIK

(med anvendelse af Excel)



1. udgave 2007

FORORD

Notatet er bygget op således, at de væsentligste begreber søges forklaret anskueligt og ved hjælp af et stort antal eksempler.
Disse eksempler er fortrinsvis regnet ved anvendelse af de indbyggede statistikfunktioner i Excel.
Det forudsættes derfor, at læseren har adgang til en PC med Excel.

I en del eksempler vil det blive forudsat, at man kan hente data fra " nettet" eksempelvis fra www.statistikbanken.dk

Der vil i ringe omfang blive benyttet statistiske tabeller.
Det kan dog være en fordel, at have adgang til en lommeregner, da man enkle beregninger ofte derved kan udføres hurtigere.

Andre notater indenfor statistik og matematik kan i pdf-format findes på adressen www.larsen-net.dk

September 2007 Mogens Oddershede Larsen.

INDHOLD

1 Deskriptiv Statistik

1.1 Indledning

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

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

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

1.2.Grafisk beskrivelse af data

I den **deskriptive statistik** (eller beskrivende statistik) beskrives de indsamlede data i form af tabeller, søjlediagrammer, lagkagediagrammer, kurver samt ved udregning af centrale tal som gennemsnit, typetal, spredning osv.
Kurver og diagrammer forstås lettere og mere umiddelbart end kolonner af tal i en tabel. Øjet er uovertruffet til mønstergenkendelse ("en tegning siger mere end 1000 ord").

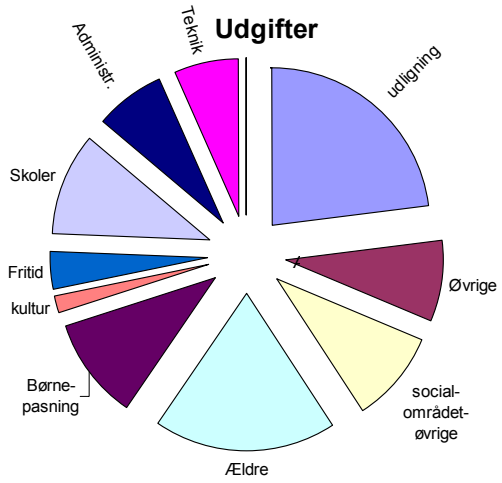
1.2.1 Kvalitative data

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.

Til illustration af disse data bruges sædvanligvis lagkagediagrammer eller søjlediagrammer

Eksempel 1.1 Lagkagediagram

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



I Excel opskrives

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

Excel-ordrer:
Vælg på værktøjslinien "Guiden diagram" ▶ Cirkel ▶ Marker ønsket figur ▶ Næste ▶ marker udskriftsområde ▶ Næste ▶ - Navn på kategori ▶ Udfør

Slet eventuelt listen med navne

◆

Eksempel 1.2 (kvalitative data)

Følgende tabel angiver mandattallet ved de to sidste folketingsvalg.

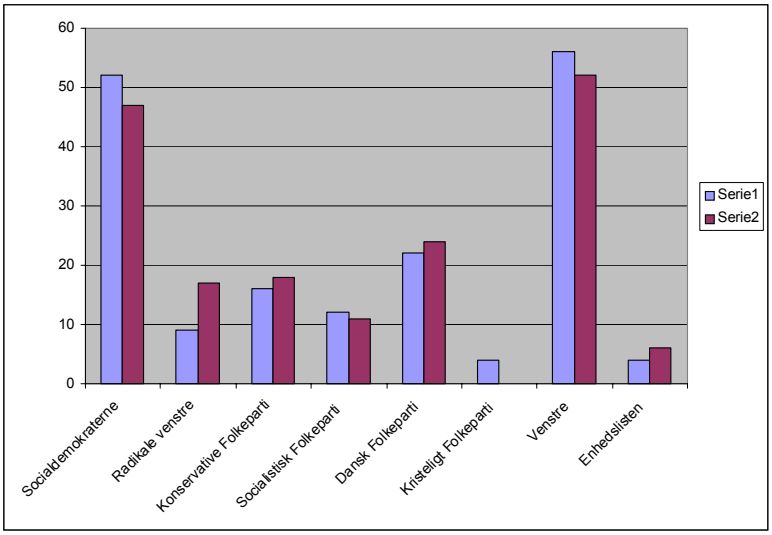
Partier		A	B	C	F	K	O	V	Ø
Mandater	2001	52	9	16	12	4	22	56	4
	2005	47	17	18	11	0	24	52	6

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

Et søjlediagram fås i Excel ved at opskrive

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

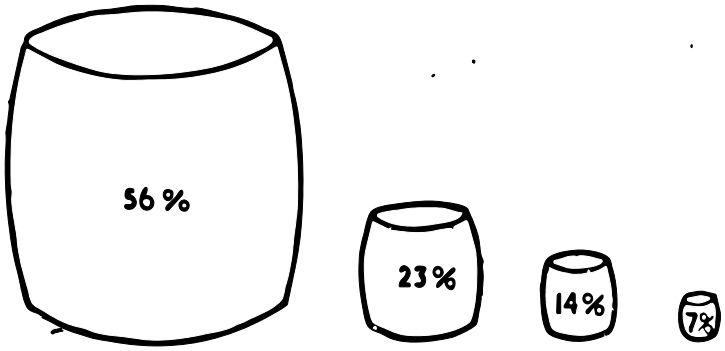
Vælg på værktøjslinien "Guiden diagram" ▶ Søjle ▶ Marker ønsket figur ▶ Næste ▶ marker udskriftsområde ▶ Næste ▶ Næste ▶ Udfør



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.
Nedenstående eksempel viser hvordan en figur kan være misvisende uden direkte at være forkert.

Eksempel 1.3. 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.

1.2.2. Kvantitative data (variable)

Kvantitative data er data, hvor registreringen i sig selv er tal, der angiver en bestemt rækkefølge, f.eks som i eksempel 1.4 hvor data regristreres efter det tidspunkt hvor registreringen foregår eller som i eksempel 1.5, hvor det er størrelsen af regristrerede værdi der er af interesse.

Eksempel 1.4. 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" ▶ Ind- og udvandring ▶ Ind- og udvandring efter bevægelse ▶ under "bevægelse" vælges alle og under "måned" vælges år og derefter alle ▶ Tryk på tabel ▶ Drej tabel med uret ▶ Gem som Excel fil

Indvandringer og udvandringer efter tid

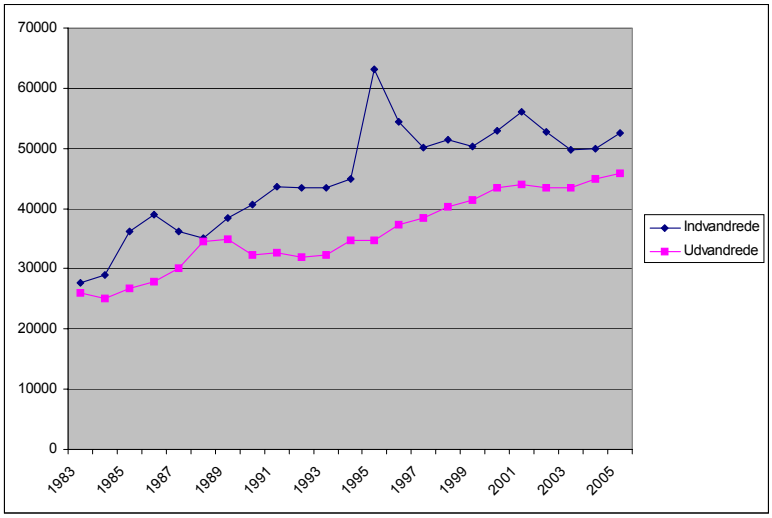
	Indvandrede	Udvandrede
1983	27718	25999
1984	29035	25053
1985	36214	26715
1986	38932	27928
1987	36296	30123
1988	35051	34544
1989	38391	34949
1990	40715	32383
1991	43567	32629
1992	43377	31915
1993	43400	32344
1994	44961	34710
1995	63187	34630
1996	54445	37312
1997	50105	38393
1998	51372	40340
1999	50236	41340
2000	52915	43417
2001	55984	43980
2002	52778	43481
2003	49754	43466
2004	49860	45017
2005	52458	45869

Giv en grafisk beskrivelse af disse data.

Løsning:

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

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



Eksempel 1.5. Kvantitativ variabel , sideafvigelse ved skydning.
Man har 100 gange målt sideafvigelsen ved skydning med maskingevær.
Resultaterne var følgende:

33.22	21.75	11.60	4.12	13.19	11.03	-0.8	-19.01	11.08	10.91	6.93	14.6
-11.5	2.19	14.47	11.27	22.06	11.81	19.53	13.25	6.1	1.14	14.1	-4.23
9.33	14.26	-4.16	20.88	-13.29	-6.53	-3.03	0.49	13.08	3.7	-0.56	-0.36
22.29	9.01	21.49	5.1	17.88	2.68	5.23	2.81	-5.64	11.63	3.21	-0.19
18.67	17.01	-6.34	21.6	11.26	9.63	-5.97	6.42	14.65	-0.77	0.31	-0.43
2.26	6.14	12.56	11.81	11.76	23.92	3.66	23.98	3.81	26.44	4.67	21.38
-0.52	5.51	-24.44	-5	13.95	-6.66	10.63	10.55	-1.69	-0.37	12.59	24.22
24.16	30.22	-11.84	14.45	-12.27	18.94	0.85	12.93	8.89	13.64	-3.28	16.27
16.63	5.87	4.35	13.7								

Giv en grafisk beskrivelse af disse data.

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

Klasser		Antal *n*
]-24.0 ; -19.7]	/	1
]-19.7 ; -13.9]	/	1
]-13.9 ; - 8.1]	////	4
]-8.1 ; - 2.3]	//////////	10
]-2.3 ; 3.5]	//////////////////	18
]3.5 ; 9.3]	/////////////////	16
]9.3 ; 15.1]	//////////////////////////////	29
]15.1 ; 20.9]	////////	8
]20.9 ; 26.7]	//////////	11
]26.7 ; 33.5]	//	2

Det ses, at de fleste målinger ligger fra ca - 1.8 til ca 15.1 og så falder hyppigheden nogenlunde symmetrisk til begge sider.

I Excel sker det på følgende måde:

Data indtastes i eksempelvis søjle A1 til A100
Vælg "Funktioner", Dataanalyse, Histogram
I den fremkomne tabel udfyldes "inputområdet" med A1:A100 og man vælger "diagramoutput"..

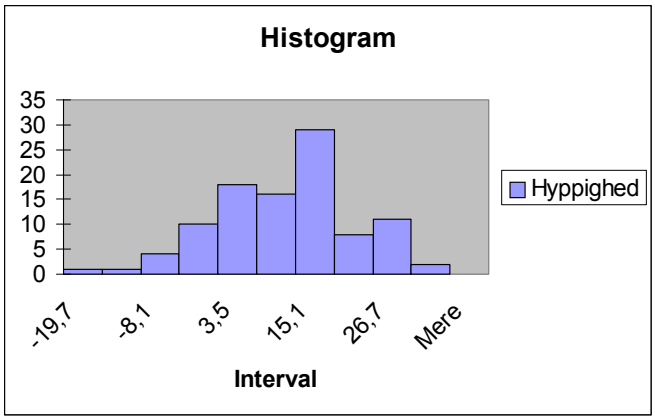
1) Trykkes på OK fås en tabel med hyppigheder, og en figur, hvor intervalgrænserne er fastlagt af Excel.
2) Ønsker man selv at bestemme grænserne, skal man også udfylde intervalområdet. Dette gøres ved at skrive de øvre grænser i en søjle (f.eks i B1 -19.7, i B2 -13.9 osv) og så skrive B1:B11 i inputområdet

Nedenstående figurer er blevet gjort lidt "pænere" ved

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

I tilfælde 2 fremkommer følgende

Interval	*Hyppighed*
-19,7	1
-13,9	1
-8,1	4
-2,3	10
3,5	18
9,3	16
15,1	29
20,9	8
26,7	11
33,5	2
Mere	0



Histogrammet viser også, at de fleste tal ligger fra -1.8 til 15.1, 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 (beskrives nærmere i næste kapitel)

Sumpolygon

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

Eksempel 1.6 Sumpolygon

Lad os igen betragte de 100 sideafvigelser i eksempel 1.5, og foretage den i den følgende tabel angivne opsummering(kumulering).
Afsættes punkterne (-19.7 , 0.01), (-13.9, 0.02) . . . (38.3, 1.00) (bemærk at x-værdierne er værdierne i højre intervalendepunkt), og forbindes de enkelte punkter med rette linier, fås den i figur 1.1 angivne sumpolygon, hvoraf man kan aflæse, at 25% af sideafvigelserne ligger under ca. -2. (kaldes 25% fraktilen eller første kvartil).

Klasser	Antal	Kumuleret relativ hyppighed i %
]-25.5 ; -19.7]	1	1
]-19.2 ; -13.9]	1	2
]-13.4 ; - 8.1]	4	6
]-7.6 ; - 2.3]	10	16
]-1.8 ; 3.5]	18	34
]4.0 ; 9.3]	16	50
]9.8 ; 15.1]	29	79
]15.6 ; 20.9]	8	87
]21.4 ; 26.7]	11	98
]27.2 ; 33.5]	2	100

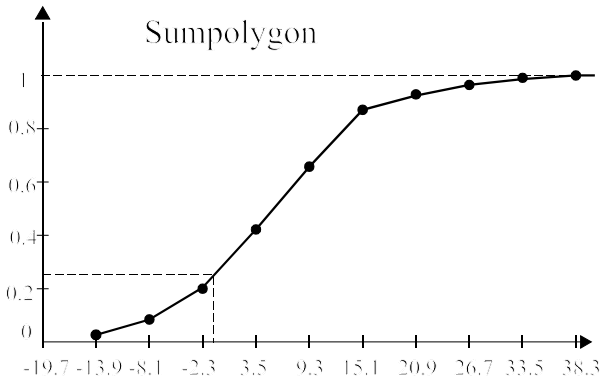
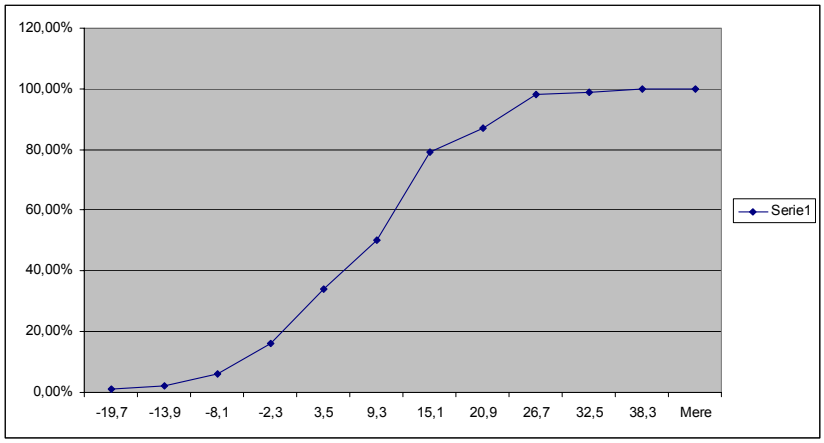


Fig 1.1 *Sumpolygon*

I Excel fås en sumpolygon på følgende måde:
Data indtastes i eksempelvis søjle A1 til A100
Vælg "Funktioner" ▶ , Dataanalyse ▶ Histogram
I den fremkomne tabel udfyldes "inputområdet" og man vælger "kumulativ frekvens".
Trykkes på OK fås en tabel med hyppigheder og kumulerede frekvenser.
Marker intervalsøjlen og komulativ søjle ▶ I værktøjslinien vælges "diagram" ▶ vælg "kurve" osv. ▶ udfør.

Interval	*Hyppighed*	*Kumulativ %*
-19,7	1	1,00%
-13,9	1	2,00%
-8,1	4	6,00%
-2,3	10	16,00%
3,5	18	34,00%
9,3	16	50,00%
15,1	29	79,00%
20,9	8	87,00%
26,7	11	98,00%
32,5	2	100,00%
Mere	0	100,00%



1.3. 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:
Det er jo ikke sikkert at man kender alle elementer i populationen. I så fald kunne man foretage en såkaldt systematisk udvælgelse, hvor man vælger 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 populatonen 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 onskes man en vurdering af fejlprocenten.
Man udtager randomiseret 5 kasser, og undersøger alle emnerne i kasserne.

1.4. Karakteristiske tal

I dette afsnit søger man at karakterisere stikprøven og dermed hele populationen ved nogle få karakteristiske tal.

Benyttes Excel på stikprøven på de 100 sideafvigelser som vi i det følgende vil kalde *x*.
Funktioner ▶ Dataanalyse ▶ Beskrivende statistik ▶ Resumestatistik
fås følgende udskrift

	Kolonne1
Middelværdi	7,7883
Standardfejl	1,074561861
Median	9,17
Tilstand	11,81
Standardafvigelse	10,74561861
Stikprøvevarians	115,4683193
Kurtosis	0,146051416
Skævhed	-0,278076146
Område	57,66
Minimum	-24,44
Maksimum	33,22
Sum	778,83
Antal	100

Nogle af disse betegnelser er en ikke korrekt statistisk oversættelse fra engelsk, så i det følgende vil andre betegnelser ofte blive anvendt.
Ikke alle i listen er i denne sammenhæng af interesse. Til gengæld savnes en beregning af de såkaldte kvartiler
I det følgende vil vi koncentrere os om

Estimat (skøn)	Betegnelse
Midterværdi	middelværdi (burde hedde gennemsnit) og median
Spredning	standardafvigelse (spredning), stikprøvevarians, kvartilafstand, standardfejl

Begreberne vil blive gennemgået i de følgende afsnit.

1.4.1. Midterværdier

Middelværdi: Kendes den teoretiske fordeling eksakt, eller hele "populationen" (målt højden på alle danske mænd) kan beregnes en "korrekt midterværdi" kaldet middelværdi μ (græsk my) eller $E(X)$ (Expected value)
Ud fra stikprøven vil en tilnærmet værdi (kaldet et **estimat**) for μ være **gennemsnittet** $\bar{x}$ (kaldt x streg).

Generel formel: $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$ hvor observationerne i en stikprøve er

$x_1, x_2, \ldots, x_n$

Eksempel: Tallene 2,4,5,9 har gennemsnittet $\bar{x} = \frac{2+4+5+9}{4} = 5$

I Excel-udskriften under middelværdi (som altså burde hedde gennemsnit) findes for de 100 x-værdier $\bar{x}$ =7.7883.

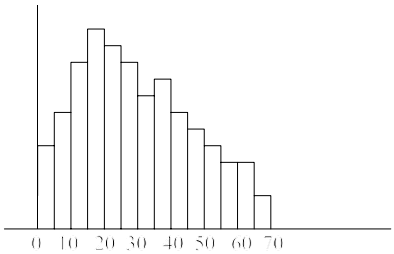
Median: Medianen beregnes på følgende måde:
1) Observationerne ordnes i rækkefølge efter størrelse.
2a) Ved et ulige antal observationer er midianen det midterste tal
2b) Ved et lige antal er medianen gennemsnittet af de to midterste tal.

Eksempel: Observationer 6, 17, 7, 13, 5, 2. Ordnet i rækkefølge: 2, 5, 6, 7 13, 17. Median 6,5

Medianen kaldes også for **50% fraktilen**, fordi den brøkdel (fraktil) der ligger under medianen er ca 50% .
Eksempelvis er medianen for de 100 x- værdier 9.17, dvs, halvdelen af sideafvigelserne ligger under 9,17.

Er median og gennemsnit nogenlunde lige store er fordelingen nogenlunde symmetrisk omkring middelværdien.

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



Er medianen større end gennemsnittet som i eksempel 1.5 med sideafvigelserne, er der muligvis tale om en venstreskæv fordeling

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

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

[1] jævnfør statistisk årbog 2005 tabel 144 eller se www.statistikbanken.dk Og vælg løn\lønstatistik for den statslige sektor\løn32\klik for at vælge\alle værdier\hovedgrupper\ledelse på højt niveau+kontorarbejde

1.4.2 Spredningsmål.

Støj

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

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

Eksempel: 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 kvartilafstand.**
Den viser også, at lønspredningen er væsentlig mindre for række 2 end for række 1 .

◆

I Excel beregnes kvartilerne således:
Data indtastes i eksempelvis søjle A1 til A100 ▶ På værktøjslinien foroven: Tryk på f_x = ▶
På rullemenu vælges "Kvartil" (evt. først vælg kategorien "statistik") ▶ Der fremkommer en tabel med anvisning på, hvordan den skal udfyldes
Med tallene fra sideafvigelserne i eksempel 1.5 fås

3 kvartil	14,3075
1. kvartil	0,185
kvartilafstand	14,1225

◆

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

Standardafvigelse (dansk: spredning, engelsk:standard deviation)
I modsætning til de forrige spredningsmål baserer standardafvigelsen sig på alle observationer i stikprøven (eller populationen) og er derfor (hvis fordelingen er nogenlunde symmetrisk (normal-fordelt) det mest anvendte mål.
Hvis spredningen baserer sig på hele populationen benævnes den $\sigma(X)$ eller kort σ .
Baserer spredningen sig kun på en stikprøve benævnes den s. Kort: s er et estimat (skøn) for σ .

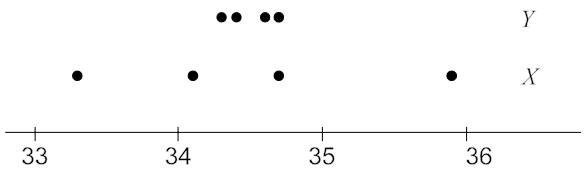
s beregnes af formlen $s = \sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$ hvor observationerne i en stikprøve er $x_1, x_2, \ldots, x_n$

Stikprøvevariansen (eller blot variansen) er s^2 .
Eksempel: Tallene 2,4,5,9 med $\bar{x} = 5$,har variansen

$$s^2 = \frac{(2-5)^2 + (4-5)^2 + (5-5)^2 + (9-5)^2}{4-1} = \frac{26}{3} = 8.666 \text{ og spredningen } s = \sqrt{8.667} = 2.94$$

Af udskriften i Excel for de 100 værdier fås s =10.7456 og $s^2 = 115{,}4683$

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.

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

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

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

Vurdering af størrelsen af stikprøvens spredning.
Man kan vise, at for tæthedsfunktioner med kun et maksimumspunkt gælder, at mellem $\bar{x} - 2 \cdot s$ og $\bar{x} + 2 \cdot s$ ligger ca. 89% af resultaterne, og mellem $\bar{x} - 3 \cdot s$ og $\bar{x} + 3 \cdot s$ ligger ca. 95% af resultaterne.

For normalfordelingen er de tilsvarende tal 95% og 99%.(se figur 1.2)

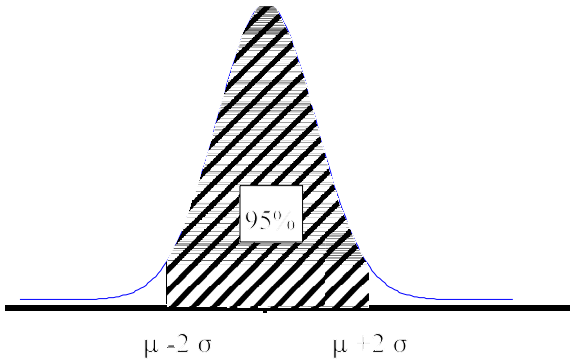


Fig 1.2. *Mellem* $\mu - 2 \cdot \sigma$ *og* $\mu + 2 \cdot \sigma$ *ligger ca.* 95% *af resultaterne.*

Dette benyttes bl.a. i statistisk kvalitetskontrol, hvor man løbende udtager stikprøver af produktionen.
Eksempelvis kan man om en måling, der giver en værdi, der ligger udenfor intervallet $[\bar{x} - 3 \cdot s; \bar{x} + 3 \cdot s]$ sige, at hvis ikke det er en fejlmåling, så er der noget galt ved produktionen (en maskine løbet varm eller lignende)

Sætning 1.1. Spredning på et gennemsnit.

Stikprøvegennemsnittet $\bar{x}$ *varierer med en spredning på* ***standardfejlen*** $s(\bar{x}) = \frac{s}{\sqrt{n}}$,

hvor n er stikprøvestørrelsen.

For tallene i eksempel 1.5 gælder således, at gennemsnittet $\bar{x} = 7.7883$ har en spredning på

$$s(\bar{x}) = \frac{10.7456}{\sqrt{100}} = 1.0746$$

Man opnår altså en væsentligt mere præcist estimat (resultat) , hvis man beregner et gennemsnit på 100 målinger, da spredningen på den enkelte måling så skal divideres med 10.
Er det meget dyre målinger er det dog sædvanligvis klogest f.eks. at nøjes med 25 målinger, og bruge ressourcerne på anden vis.
Fordelen ved at gå fra 25 målinger til 100 målinger er begrænset, da spredningen jo kun bliver halveret derved.

Variationskoefficient $= \frac{s}{\bar{x}}$.

Skal man sammenligne spredningen af to forskellige fordelinger, f.eks. spredningen af lønningerne med 10 års mellemrum, hvor der måske er sket en generel lønstigning, så kan man ikke direkte bruge de to s-værdier. Det skyldes, at hvis alle tal eksempelvis bliver fordoblet, så vil også s blive fordoblet. Man må derfor i stedet bruge variationskoefficienten, hvor man har divideret med gennemsnittet. (svarer til den relative kvartilafstand, hvor man dividerede med medianen.

Som nævnt er de øvrige tal i Excel- listen uden større interesse for os, men her gives en kort beskrivelse af dem.

Tilstand (dansk: typetal, engelsk: mode) er det tal (her 11,81) der forekommer flest gange i datasættet. Dette tal er kun i særlige tilfælde nyttigt at kende

Område (dansk"variationsbredden, engelsk: range) er afstanden mellem største og mindste talværdi. I udskriften angav Excel den til 57.66 (33.22 - (-24.44))
Den angiver kun et groft mål for, hvor meget tallene spreder sig, da den kun er baseret på to tal, og ikke inddrager alle observationerne..

Kurtosis: Angiver i hvilken grad fordelingen er spids eller flad i sammenligning med en normalfordeling. Et tal mellem -1 og 1 angiver, at der nogenlunde er tale om en normalfordeling.
Værdien 0.146 antyder at fordelingen ikke på det punkt afviger meget fra en normalfordeling.

Skævhed: Angiver et mål for hvor "skæv" fordelingen er.
Groft taget kan man sige , at en værdi under -1 angiver en kritisk skæv fordeling med toppunkt mod venstre, mens tilsvarende en positiv værdi over 1 angiver en kritisk skæv fordeling mod højre. I sådanne tilfælde bør man anvende median og ikke gennemsnit som mål.
Da skævheden i eksemplet kun er - 0.278 er skævheden ikke kritisk.

1.5. Grupperede fordelinger.

I mange statistiske tabeller angiver man for overskuelighedens skyld ikke de oprindelige data, men grupperer tallene og angiver så kun hyppighederne indenfor hver gruppe.
Excel kan ikke her automatisk beregne de forskellige karakteristiske tal, så det må gøres manuelt.
Lad os igen betragte tallene fra eksempel 1.3, men nu tænke os, at vi kun kender hyppighederne.
For at få estimat for gennemsnit og spredning antager man nu, at alle observationer ligger i midten af intervallet. Se tavlen.

Klasser	Midtpunkt x_i	Antal n	$x_i \cdot \frac{n}{100}$
]-25.5 ; -19.7]	- 22.6	1	-0.226
]-19.7 ; -13.9]	- 16.8	1	-0.168
]-13.9 ; - 8.1]	- 11.0	4	-0.44
]-8.1 ; - 2.3]	- 5.2	10	-0.52
]-2.3 ; 3.5]	0.6	18	0.108
]3.5 ; 9.3]	6.4	16	1.024
]9.3 ; 15.1]	12.2	29	3.538
]15.1 ; 20.9]	18.0	8	1.44
]20.9 ; 26.7]	23.8	11	2.618
]26.7 ; 33.5]	29.6	1	0.296
]33.5 ; 38.3]	35.4	1	0.354
SUM			8.024

Som det ses er $\bar{x}$ = 8.024 tæt ved den "korrekte" værdi 7.788 som Excel fandt.
Man siger, at gennemsnittet er et **vægtet gennemsnit**, fordi hver værdi indgår med en vægt svarende til andelen af værdier i hvert interval.

På tilsvarende måde kan man finde spredningen af formlen $\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2\cdot\frac{n}{100}}{100-1}}$

Excel har desværre ikke et egentligt program hertil.
Excel: I Excel kan man generere tilfældige tal baseret på bl.a. en "jævn" fordeling, (kaldes den rektangulære fordeling), dvs. hvor alle tal har samme sandsynlighed indenfor et angivet interval [ab].
Vi lader nu Excel danne 1 af sådanne tal i intervallet [-25.5 ; -19.7], 1 af sådanne tal i intervallet [-19.2 ; -13.9] osv.
Tallene placeres i samme søjle under hinanden.
Funktioner ▶ Dataanalyse ▶ Generering af tilfældige tal
I menu vælges"Antal variable = 1, Antal tilfældige tal = 1, Fordeling= jævn, Mellem -25.5 og -19.7 Outputareal = A1 , OK
Vi gentager nu
Funktioner ▶ Dataanalyse ▶ Generering af tilfældige tal
I menu vælges"Antal variable = 1, Antal tilfældige tal = 1, Fordeling= jævn, Mellem -19.7 og -13.9 Outputareal = A2 , OK
Således fortsættes, til vi har alle 100 tal placeret i A-søjlen.
Vi har nu nogle (ganske vist) kunstige tal, som vi kan fortsætte med at tegne histogram, sumpolygon finde karakteristiske tal osv. som før.

Ønsker man kun at beregne gennemsnit, spredning og median, kan man i mange tilfælde gøre det uden at lave "kunstige" observationer jævnfør følgende eksempel.

Eksempel 1.6 Grupperet fordeling
I statistikbanken findes følgende tabel over aldersfordelingen af elever fra København, som er under uddannelse til forsvaret i 2004.

alder	21	22	23	24	25	26	27	28	29	30-34	35-39
antal	1	11	44	48	45	34	21	22	15	19	2

Beregn gennemsnit og spredning
Løsning:
Beregningerne er foretaget i Excel

$$\bar{x}=\frac{1\cdot 21+11\cdot 22+44\cdot 23+\ldots+15\cdot 29+19\cdot 32+2\cdot 37}{1+11+44+48+45+34+21+22+15+19+2}=\frac{6736}{262}=\underline{\underline{25.7}}$$

$$s=\sqrt{\frac{1\cdot(21-25.7)^2+11\cdot(22-25.7)^2+\ldots+2\cdot(37-25.7)^2}{262-1}}=\underline{\underline{10.74}}$$

Median: Der summeres op til man når 131. Heraf ses, at median er $\underline{\underline{25}}$

Opgaver

Opgave 1.1.
Angiv i hvert af følgende tilfælde, om de følgende variable er kvantitative eller kvalitative.
a) Køn (mand eller kvinde)
b) Temperatur
c) Antal dage i den sidste uge hvor en kadet på søofficerskolen har fået mindst en drink
d) CPR-nummer

Opgave 1.2.
I www.statistikbanken.dk/luft4 er følgende oplysninger for året 2003 hentet ind i Excel.

Udslip til luft af drivhusgasser efter enhed, type, kilde og tid

			2003
Mia. C02-ækvivalenter	I alt	Energisektoren	32
		Industri og produktion	8
		Transport	13
		Affaldsbehandling	2
		Landbrug	10
		Andet	9

a) Hent selv disse data ind i Excel, og opstil et lagkagediagram til belysning af tallene.
b) Find de tilsvarende tal for 1996, og vælg en passende grafisk fremstilling til sammenligning af tallene fra 1996 og 2003.
c) Beregn i Excel for årene 1990 til 2003 energisektorens udslip i forhold til det samlede udslip af drivhusgasser (i %), og tegn dette grafisk.

Opgave 1.3
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 1.4
I http://www.statistikbanken.dk/statbank5a/default.asp?w=1600 findes nogle oplysninger om Danmarks forbrug af energi efter type og mængde.
1) Hent produktion af naturgas og råolie ind målt i tons for de sidste 2 år (i måneder) ind i Excel
2) Tegn i Excel i samme koordinatsystem to kurver for henholdsvis produktionen af naturgas og råolie.

Opgave 1.5

Færdselspolitiet overvejede, om der burde indføres en fartgrænse på 70 km/h på en bestemt landevejsstrækning, hvor der hidtil havde været en fartgrænse på 80 km/h.
Som et led i analysen af hensigtmæssigheden af den overvejede ændring observeredes inden for et bestemt tidsrum ved hjælp af radarkontrol de forbipasserende bilers fart. Resultatet af målingerne var:

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

1) Foretag en vurdering af, om fordelingen er nogenlunde symmetrisk (normalfordelt) ved
   a) at tegne et histogram
   b) at beregne karakteristiske værdier
2) Tegn en sumpolygon for fordelingen, og benyt den til at angive hvor stor en procent af bilisterne, der "approksimativt" overstiger hastighedsgrænsen på 80 km/h. (Vink: Vælg hensigtsmæssige intervalgrænser).

Opgave 1.6

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

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

1) Foretag en vurdering af, om fordelingen er nogenlunde symmetrisk (normalfordelt) ved
   a) at tegne et histogram
   b) at beregne 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 1.7

Den følgende tabel 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	2.90	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	2.90	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	2.90	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
   (Vink: Anvend histogram og kumulativ frekvens).

Opgave 1.8

I "statistikbanken" http://www.statistikbanken.dk/statbank5a/default.asp?w=1600 finder man under punktet "Uddannelse og kultur","elever pr. 1 oktober", U11: Elever efter bopælskommune osv, en statistik over antal elever i forsvaret (se under punkt 5095) i 2004 fordelt efter alder for hele landet.

1) Indsæt data i Excel for såvel mænd som kvinder

2a) Lav et søjlediagram over aldersfordelingen for mænd. Bemærk, at da intervallerne ikke er lige lange, må man ændre på inddelingerne.

2b) Beregn median , middelværdi, spredning og relativ kvartilafstand for mænd.
Vurder om fordelingen er symmetrisk, venstreskæv eller højreskæv..
Vink: Da fordelingen er grupperet, må man i Excel lave "kunstige" data.

3) Lav de til spørgsmål 2a) og 2b) svarende beregninger for kvinder.

Opgave 1.9

I http://www.statistikbanken.dk/statbank5a/default.asp?w=1600 findes under Løn og lønstatistik for statslige ansatte under "løn 31" nogle oplysninger om fortjenesten for statsansatte efter uddannelse m.m. i forsvar i 2005.

1) Angiv gennemsnit, median, øvre og nedre kvartil for såvel mænd som kvinder
2) Overfør data til Excel på egen harddisk
3) Angiv om de to fordelinger er symmetrisk, højre eller venstreskæv
4) Er der forskel på lønspredningen for mænd og kvinder
   (Vink: Beregn den relative kvartilafstand)

2 NORMALFORDELINGEN

2.1 Indledning

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

2.2. Relative hyppigheder , tæthedsfunktion.

2.2.1.Relative hyppigheder.

Ved den relative hyppighed forstås hyppigheden divideret med det totale antal.
I eksempel 1.5 er den relative hyppighed for sideafvigelsen i intervallet]4.0 ; 9.3] $\frac{16}{100} = 16\%$
Man kunne sige, at "sandsynligheden" er 16% for at sideafvigelsen ligger i dette interval.
Skal man sammenligne to talmaterialer, eksempelvis sammenligne de 100-værdier i eksempel 1.5 med 200 resultater fra en anden skydebane, har det ingen mening at sammenligne hyppighederne, men derimod de relative hyppigheder, dvs. dividere hyppighederne med henholdsvis 100 og 200.

2.2.2. Tæthedsfunktion.

Efter at man har indsamlet data, vil man søge ud fra stikprøven at få et indtryk af karakteristiske træk ved hele populationen. Her spiller tæthedsfunktionen en vigtig rolle.
Vi vil igen benytte eksempel 1.5 til at anskueliggøre denne funktion

Eksempel 2.1 . Tæthedsfunktion

I den følgende tabel er dels beregnet de relative hyppigheder for tallene i eksempel 1.5 dels er der af hensyn til det følgende foretaget en skalering ved at dividere den relative hyppighed med intervallængden 5.8.

Klasser	Antal n	Relativ hyppighed $\frac{n}{100}$	Skalering $\frac{n}{100 \cdot 5.8}$
]-25.5 ; -19.7]	1	0.01	0.001724
]-19.7 ; -13.9]	1	0.01	0.001724
]-13.9 ; - 8.1]	4	0.04	0.006897
]-8.1 ; - 2.3]	10	0.1	0.017241
]-2.3 ; 3.5]	18	0.18	0.031034
]3.5 ; 9.3]	16	0.16	0.027586
]9.3 ; 15.1]	29	0.29	0.050
]15.1 ; 20.9]	8	0.08	0.013793
]20.9 ; 26.7]	11	0.11	0.018966
]26.7 ; 33.5]	1	0.01	0.001724
]33.5 ; 38.3]	1	0.01	0.001724

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

◆

Eksempel 2.1 begrunder, at en tæthedsfunktion for en kontinuert stokastisk variabel X skal have den egenskab, at sandsynligheden for at X ligger mellem 2 værdier a og b lig med arealet under kurven[1].

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

(P står for probability)

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

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

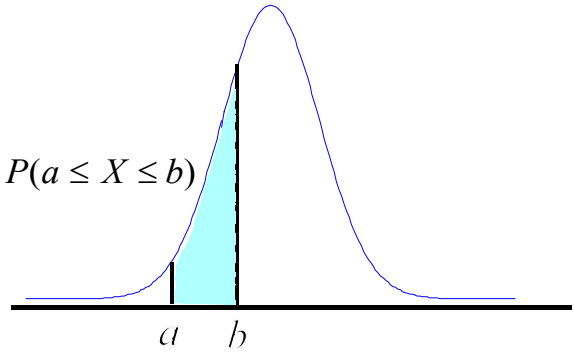


Fig 2.1 *Tæthedsfunktion*

Middelværdi:
På basis af en stikprøve på n tal , kunne vi regne gennemsnit $\bar{x}$ og spredning s ud.
Kendes den stokastiske variabel X's tæthedsfunktion $f(x)$ kan beregnes en "korrekt midterværdi".
Denne kaldes middelværdien for X og benævnes μ eller $E(X)$ (E for expected).[2]

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

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

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

Fordelingsfunktion

Svarende til at vi tegnede sumkurven i eksempel 1.6 for en stikprøve, kan vi tilsvarende definere en såkaldr **fordelingsfunktionen** $F(x)$ for en stokastisk variabel X.
Denne er defineret[3] ved $F(x) = P(X \le x)$

Grafen for $F(x)$ kan ses på figur 2.1

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

Medianen m er 50% fraktilen.

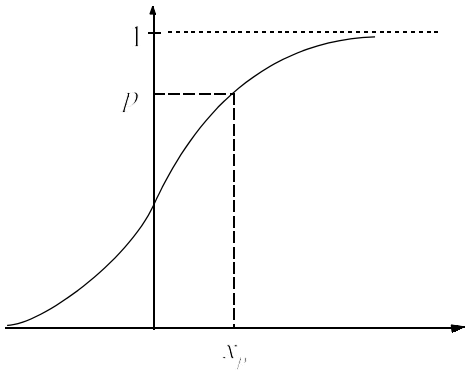


Fig 2.1 *Fordelingsfunktion*

---

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

Varians $\mathrm{V(X)} = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)dx$

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

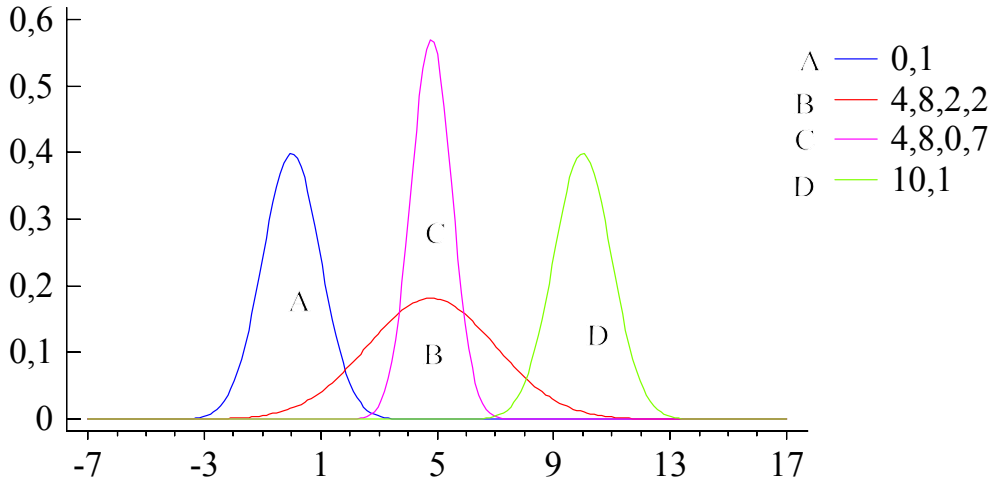
[3] $F(x) = P(X \le x) = \int_{-\infty}^{x} f(x)dx$

2.3 Definition og beregning .

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

Tæthedsfunktionen $f(x)$ [4] og den tilsvarende fordelingsfunktion findes på Excel og mange "matematiklommeregnere. Tidligere benyttede man også tabeller over den

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 $n(10\ ,\ 1)$.



Arealerne under kurverne er alle 1, og man ser, at "klokkeformen" bliver bred når spredningen er stor. Som tidligere nævnt vil et interval på $[\mu - 3\cdot\sigma;\mu + 3\cdot\sigma]$ indeholder stort set hele sandsynlighedsmassen.

Beregning af sandsynligheder

Excel:

På værktøjslinien foroven: Tryk f_x ▶ Vælg kategorien "Statistisk" ▶ Vælg "NORMALFORDELING" eller NormINV.

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

$P(X < x)$ =NORMFORDELING(x; μ ; σ ;1)
$P(X \le x_p) = p \quad x_p$ = NORMINV(p; μ ; σ)

Følgende eksempel illustrerer hvordan man i Excel beregner sandsynligheder i normalfordelingen.

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

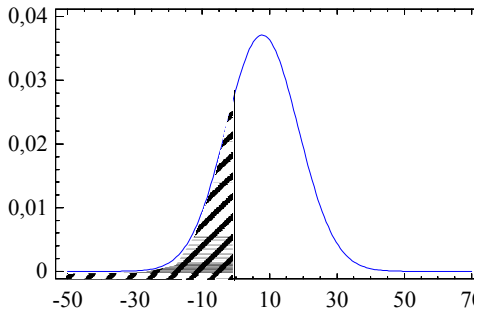
Eksempel 2.2 Beregning af sandsynligheder i normalfordelingen

For den i eksempel 1.5 angivne stikprøve på sideafvigelserne ved 100 skud fandt vi at gennemsnittet var $\bar{x}$ =7.79 og spredningen s = 10.75.
Vi antager nu, at den stokastiske variabel X = sideafvigelserne ved affyring med maskingevær er med tilnærmelse normalfordelt $n(\mu,\sigma)$ med en middelværdi $\mu = 7.79$ og en spredning på $\sigma = 10.75$.

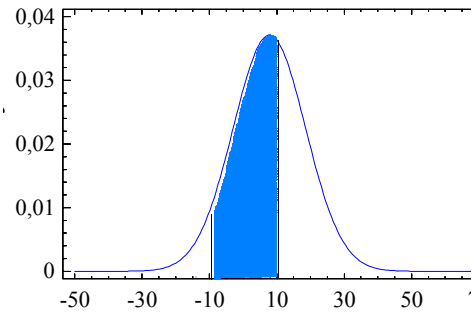
1) Beregn sandsynligheden for, at skudene rammer til venstre for målskiven, dvs X er mindre end 0
2) Beregn sandsynligheden for at skuddene falder i en afstand fra målskiven, som er mindre end 10, dvs at X ligger mellem -10.0 og 10.0
3) Beregn sandsynligheden for, at skuddene rammer til højre for målskiven, i en afstand, der er større end 15, dvs. at X er større end 15.
4) Beregn medianen m, dvs. find den sideafvigelse m som halvdelen af skuddene ligger til venstre for. Dette er det samme som at finde m, så $P(X \le m) = 0.50$
5) Beregn 95% fraktilen, dvs. den værdi $x_{0.95}$, som 95% af sideafvigelserne ligger til venstre for. Det er det samme som at sige, at $P(X \le x_{0.95}) = 0.95$

Løsning:

1) Sandsynligheden for, at sideafvigelserne er mindre end 0, er lig med arealet af det skraverede område under tæthedsfunktionen (se figuren).
   Resultat:
   $P(X \le 0)$ = NORMFORDELING(0;7,79;10,75;1)= 0,234333 ≈ 23.4%



2) Sandsynligheden for, at sideafvigelserne ligger mellem -10 og 10 er lig med arealet af det farvede område under tæthedsfunktionen (se figuren).



   Beregningen sker i Excel ved at beregne arealet fra $-\infty$ til 10 og derfra trække arealet fra $-\infty$ til -10, dvs.
   $P(-10 \le X \le 10) = P(X \le 10) - P(X \le -10)$ =
   NORMFORDELING(10;7,79;10,75;1)- NORMFORDELING(-10;7,79;10,75;1) = 53.25%

3) Da arealet under kurven er 1, fås
   $P(X \ge 15) = 1 - P(X < 15)$ =1 - NORMFORDELING(15;7,79;10,75;1)= 0,251207 = 25.12%

4) Her kendes arealet = 0.5 og man skal finde den tilsvarende x - værdi, dvs. vi ser på den omvendte funktion og beregner medianen m = NORMINV(0,5;7,79;10,75) = 7,79

5) Tilsvarende fås $x_{0.95}$ = NORMINV(0,95;7,79;10,75) = 25,472 ◆

Beregning ved tabel.

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

Ved beregningerne er det ofte nødvendigt at anvende følgende sammenhæng mellem fraktiler for X og fraktiler for U:

$$x_p = u_p \cdot \sigma + \mu$$

I sådanne tilfælde kan det være hurtigere at benytte en tabel over ofte benyttede værdier af U-fordelingens p - fraktiler u_p.. Disse er derfor angivet i tabel 1 sidst i notatet.

Vi vil se ovennævnte relation benyttet i det følgende eksempel.

Eksempel 2.3. Normalfordeling.

En fabrik støber plastikkasser. Fabrikken får en ordre på kasser, som blandt andet har den specifikation, at kasserne skal have en længde på 90 cm. Kasser, hvis længder ikke ligger mellem 89.2 og 90.8 cm bliver kasseret.
Det vides, at fabrikken producerer kasserne med en længde X, som er normalfordelt med en spredning på 0.5 cm.

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

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

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

En ny maskine indkøbes, og som et led i en undersøgelse af, om der dermed er sket ændringer i middelværdi og spredning produceres 12 kasser ved anvendelse af denne maskine.
Man fandt følgende længder: 89.2 90.2 89.4 90.0 90.3 89.7 89.6 89.9 90.5 90.3 89.9 90.6.

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

---

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

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

Løsning:

1) $P(89.2 < X \le 90.8) = P(X \le 90.8) - P(X \le 89.2)$
=NORMFORDELING(90,8;89,6;0,5;SAND)-NORMFORDELING(89,2;89,6;0,5;SAND)= 0,779947 ≈ 78.0%

2) Middelværdien må nu sættes til midtpunktet af intervallet, dvs. til 90 cm.
$P(X > 90.8) + P(X < 89.2) = 1 - P(X \le 90.8) + P(X < 89.2)$
=1-NORMFORDELING(90,8;90;0,5;SAND)+NORMFORDELING(89,2;90;0,5;SAND)= 0,109599 ≈ 10.96%

3) $P(89.2 < X < 90.8) = 0.95 \Leftrightarrow P(X \le 89.2) = 0.025$ (da der ligger 5% udenfor intervallet, og af symmetrigrunde må så 2,5% ligge på hver sin side af intervallet.)

Metode 1:Ved indsættelse i ligningen $x_p = u_p \cdot \sigma + \mu$ fås nu $89.2 = u_{0.025} \cdot \sigma + 90 \Leftrightarrow \sigma = \dfrac{89.2 - 90}{u_{0.025}}$

Benyttes tabel 1 fås $u_{0.025} = -1.96$ og dermed $\sigma = 0.408$

Benyttes Excel fås σ =(89,2-90)/NORMINV(0,025;0;1)=0,408171 ≈ 0.408

Metode 2: I celle A1 skrives en startværdi for σ eksempelvis 0,5.
▶ I celle B1 skrives =NORMFORDELING(89,2;90;A1;SAND) ▶ Funktioner ▶ "Målsøgning"
I "Angiv celle" skrives B1. I "Til Værdi" skrives 0,025. I "Ved ændring af celle" skrives A1.

Facit :0,408444

4) Med samme begrundelse som under punkt 3 fås:
$P(90.0 - d < X < 90.0 + d) = 0.95 \Leftrightarrow P(X \le 90.0 - d) = 0.025$ og $P(X \le 90.0 + d) = 0.975$.
Vi får nedre grænse =NORMINV(0,025;90;0,5) = 89,02002 = 89.0
Øvre grænse =NORMINV(0,975;90;0,5) = 90,97998 = 91.0

5) Ved indtastning af de 12 tal i Excel i cellerne A1 til A12 findes $\bar{x}$ = Middel(A1: A12) = 89.97
og s = STDAFV(A1:A2) = 0.435

◆

2.4. Usikkerhedsberegning

Årsagen til at gentagne målinger af en størrelse giver lidt forskelligt resultat fra geng til gang kan dels skyldes måleobjektet, dels målemetoden.
Eksempel: Er formålet at måle højden på alle rekrutter af årgang 2006, vil de enkelte højder naturligvis svinge meget, da "måleobjektet = rekrutterne" jo har forskellig højde. Usikkerheden afhænger derfor af måleobjektet, hvorimod målemetoden ikke vil have nogen væsentlig betydning for usikkerheden.
Er formålet derimod at måle en bestemt rekruts højde, så vil usikkerheden heraf afhænge af hvilken målemetode der anvendes.
Følgende tommelfinger regel gælder derfor:
En målemetodes usikkerhed findes ved at anvende metoden på et tilstrækkeligt veldefineret objekt.
Et objekts usikkerhed findes ved at måle det med en tilstrækkelig sikker målemetode.

Vi vil i dette afsnit fortrinsvis se på usikkerheder ved målemetoden.

Ved begrebet tolerance forstås den største tilladte afvigelse fra en tilstræbt værdi af en størrelse.

Eksempel: Ved seriefremstilling af elektriske modstande på 100 ohm kunne tolerancen være $1^0/_{00}$, dvs. modstandene skal ligge indenfor 100 ± 0.1 ohm.
Da vi ved, at for en normalfordelt variabel ligger ca 95% af fordelingen indenfor $2 \cdot \sigma$ og over 99% ligger indenfor $3 \cdot \sigma$, er det jo i sådan et tilfælde rimeligt at sige, at spredningen på måleresultaterne bør holde sig under eksempelvis $\frac{0.1}{3} = 0.033$ ohm.

Det skal her bemærkes, at der bør skelnes mellem "maksimal fejl", hvilket tolerancen jo er et udtryk for, og "statistisk fejl", som spredningen jo er et udtryk for.

Usikkerhed på en sammensat måling.
Det følgende eksempel illustrer hvad der menes med en "sammensat måling"

Eksempel 2.4 Usikkerhedsberegning

Måles trykket P, volumenet V og temperaturen T af en ideal gas, optræder der tilfældige målefejl, som gør værdierne usikre. Beregnes molantallet n nu af ligningen $P \cdot V = n \cdot R \cdot T$, bliver værdien af n derfor også usikker.
Vi ønsker at kunne beregne spredningen på n ud fra spredningerne på P, V og T. ◆

Der gælder en generel regel "Ophobningsloven"[1], men vi vil her kun se på et par ofte forekommende specialtilfælde.
Lad X, Y og Z være statistiske uafhængige variable med spredningerne σ_x, σ_y og σ_z
I praksis vil man i nedenstående regler erstattes σ_x med s_x osv.

Regel 1 (sumregel): *Lad* $W = a + b \cdot X + c \cdot Y + d \cdot Z$.

Der gælder da, at $\sigma(Y) = \sqrt{(b \cdot \sigma_x)^2 + (c \cdot \sigma_y)^2 + (d \cdot \sigma_z)^2}$

Regel 2(produkt regel): *Lad* $W = k \cdot X^a \cdot Y^b \cdot Z^c$

Der gælder da: $\frac{\sigma(W)}{W} = \sqrt{\left(a \cdot \frac{\sigma_x}{X}\right)^2 + \left(b \cdot \frac{\sigma_y}{Y}\right)^2 + \left(c \cdot \frac{\sigma_z}{Z}\right)^2}$

Ved den relative usikkerhed på W forstås $\frac{\sigma(W)}{W}$

[1] **Ophobningsloven for usikkerheder:** Lad $X_1, X_2, ..., X_k$ være **statistisk uafhængige** variable, som hidrører fra målinger, med usikkerhederne $\sigma_1, \sigma_2, \ldots, \sigma_k$. Beregnes en ny størrelse $Y(X_1, X_2, \ldots, X_k)$, *får Y også en usikkerhed* $\sigma(Y)$, *som kan beregnes tilnærmet af* $\sigma(Y) \approx \sqrt{\left(\frac{\partial Y}{\partial X_1} \cdot \sigma_1\right)^2 + \left(\frac{\partial Y}{\partial X_2} \cdot \sigma_2\right)^2 + \ldots + \left(\frac{\partial Y}{\partial X_k} \cdot \sigma_k\right)^2}$

Eksempel 2.5 Sumregel.
Insektpulver sælges i papkartoner. I middel fyldes der 500 gram insektpulver i hver karton med en usikkerhed (spredning) på 5 gram. Kartonen vejer i middel 10 gram med en usikkerhed (spredning) på 1.0 gram. Det kan antages, at de to vægte er statistisk uafhængige
Find bruttovægten og usikkerheden på bruttovægten
Løsning:
Lad X være vægten af insektpulveret, og lad Y være vægten af kartonen.
Bruttovægten er da Z=X+Y
Vi har Z = 500 +10 = 510 gram

Usikkerheden på Z er $\sigma(Z) = \sqrt{5^2 + 1^2} = \sqrt{26} \approx 5.10$

Bemærk: Skrivemåden $X = 500 \pm 5$ er uklar.
Sommetider menes, spredningen, og andre gange menes der maksimal fejl.
Hvis det var den maksimale fejl, der var henholdsvis 5 og 1 gram ville den maksimale fejl på Z naturligvis være 5 +1 = 6 gram.

Eksempel 2.6. Produktregel (fortsættelse af eksempel 2.4)
En ideal gas opfylder ligningen $P \cdot V = n \cdot R \cdot T$, hvor $R = 8.314\ \mathrm{J \cdot K^{-1} \cdot mol^{-1}}$.
Man har målt $P = 123400\ \mathrm{Pa}$, $V = 5.67\ \mathrm{m^3}$, $T = 678\ \mathrm{K}$
med usikkerheder $\sigma(P) = 1000\ \mathrm{Pa}$, $\sigma(V) = 0.06\ \mathrm{m^3}$, $\sigma(T) = 3\ \mathrm{K}$.
Det kan antages, at måleresultaterne for P, V og T er statistisk uafhængige.
Find molantallet n, usikkerheden $\sigma(n)$, samt den relative usikkerhed $\mathrm{rel}(n) = \dfrac{\sigma(n)}{n}$.
Løsning:

Vi finder $n = \dfrac{P \cdot V}{R \cdot T} = \dfrac{123400 \cdot 5.67}{8.314 \cdot 678} = 124.12\ \mathrm{mol}$

Vi foretager følgende omformning: $n = \dfrac{P \cdot V}{R \cdot T} = \dfrac{1}{R} \cdot P^1 \cdot V^1 \cdot T^{-1}$

Vi ser nu at produktreglen kan anvendes.

$$\frac{\sigma(n)}{n} = \sqrt{\left(a \cdot \frac{\sigma_P}{P}\right)^2 + \left(b \cdot \frac{\sigma_V}{V}\right)^2 + \left(c \cdot \frac{\sigma_T}{T}\right)^2} = \sqrt{\left(1 \cdot \frac{1000}{123400}\right)^2 + \left(1 \cdot \frac{0.06}{5.67}\right)^2 + \left((-1) \cdot \frac{3}{678}\right)^2}$$

$= 0.014044 = 1.40\%$

$\sigma(n) = 0.014044 \cdot 124.12 = 1.743\ \mathrm{mol}$

OPGAVER

Opgave 2.1

1) En statistisk 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 statistisk variabel X er normalfordelt med $\mu = 25.1$ og $\sigma = 2.4$.
   Find $P(22.3 < X \le 27.8)$.

Opgave 2.2

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

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

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

4) Giv et estimat for middelværdien og spredningen.

Opgave 2.3

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 2.4

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

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

Opgave 2.5
En mængde råmateriale til en produktion ligger i kegleformet bunke. En kegle med radius R og højde H har volumenet $V = \frac{\pi}{3} R^2 H$.
Man har målt $R = 12.0$ m , $H = 11.0$ m ,
med usikkerheder $\sigma(R) = 0.2$ m , $\sigma(H) = 0.1$ m .
Det kan antages, at måleresultaterne for R og H er statistisk uafhængige.
Find volumenet V, usikkerheden $\sigma(V)$, samt den relative usikkerhed $rel(V)$.

Opgave 2.6
For en rektangulær flade har man målt længden L og bredden B *:*
$L = 12.3$ m , $B = 8.4$ m
med usikkerheder
$\sigma(L) = 0.1$ m , $\sigma(B) = 0.2$ m .
Det kan antages, at måleresultaterne for L og B er statistisk uafhængige.
Find fladens areal A, usikkerheden $\sigma(A)$, samt den relative usikkerhed $rel(A)$.

Opgave 2.7
For et bassin af form som en retvinklet kasse har man målt længden L *,* bredden B og højden H *:*
$L = 18.0$ m , $B = 12.3$ m $H = 4.5$ m
med usikkerheder
$\sigma(L) = 0.2$ m , $\sigma(B) = 0.1$ m , $\sigma(H) = 0.2$ m .
Det kan antages, at måleresultaterne for L, B og H er statistisk uafhængige.
Find bassinets volumen V, usikkerheden $\sigma(V)$, samt den relative usikkerhed $rel(V)$.

Opgave 2.8
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.
1) Beregn sandsynligheden p for, at en tilfældig pakke opvaskemiddel har en bruttovægt mellem 4250 g og 4300 g.
2) Find usikkerheden på bruttovægten

3. Konfidensinterval

3.1. Indledning

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 hvor meget sideafvisningen var for det pågældende maskingevær. Vi fandt, at for gennemsnittet af 100 skud var den 7.79 enheder mod venstre.
Et sådant gennemsnit er imidlertid også behæftet med en vis usikkerhed.
Havde vi skudt andre 100 skud, havde vi uden tvivl fået et lidt andet gennemsnit.
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.

3.2. Fordeling og spredning af gennemsnit

Den centrale grænseværdisætning:
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 1.1 , 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 3.1. 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, at $P(7 < \overline{X} < 9) = P(X < 9) - P(X < 7) =$
NORMFORDELING(9;8,2;0,424;1)-NORMFORDELING(7;8,2;0,429;1) =0,9681 = 96.8%◆

3.3. Konfidensinterval for middelværdi

3.2.1. 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 figur 3.1, så er der mindre end 2.5% chance for, at vi ville have fået det fundne gennemsnit $\bar{x}$.

For at finde grænsen for intervallet, må vi finde en middelværdi μ så $P(\bar{X} \le \bar{x}) = 0.025$.

Lad os illustrere det ved følgende eksempel:

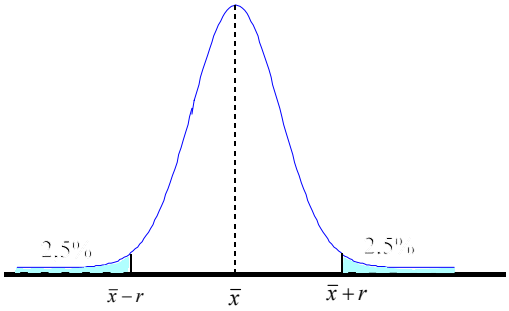


Fig 3.1. *95% konfidensinterval*

Eksempel 3.2 Beregning af 95% konfidensinterval

Lad gennemsnittet af 12 målinger være $\bar{x} = 90$

Lad os antages at spredningen kendes eksakt til $\sigma = 0.5$.

Vi ved, at spredningen på gennemsnittet er "standardfejlen" $\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{0.5}{\sqrt{12}}$.

Hvis den sande middelværdi μ afviger stærkt fra 90 er det yderst usandsynligt, at vi ville have fået et gennemsnittet på 90.

Eksempelvis, hvis $\mu = 92$ bliver

$P(\bar{X} \le 90) = \text{NORMFORDELING}(90;92;0,5/\text{KVROD}(12);1) = 0$

dvs. det er ganske usandsynligt at den sande middelværdi var 92.

For at finde grænsen kunne man finde μ af ligningen $P(\bar{X} \le 90) = 0.025$ dvs. finde μ af NORMFORDELING(90; μ;0,5/KVROD(12);1) = 0.025 [1]

Lettere er det at benytte formlen $x_p = \mu + u_p \cdot \sigma$ som ved benyttelse af, at $\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ giver

$\mu = \bar{x} - u_{0.025} \cdot \frac{\sigma}{\sqrt{12}}$. Indsættes fra tabel 1 $u_{0.025} = -1.96$ (eller =NORMINV(0,025;0;1)) fås, at

øvre grænse for konfidensintervallet er $\mu = \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{12}} = 90 + 1.96 \cdot \frac{0.5}{\sqrt{12}} = 90.283$.

Da der er symmetri omkring $\bar{x}$ fås konfidensintervallet [89.717 ; 90.283] ◆

Som det fremgår af eksempel 3.2 gælder følgende

---

[1] I celle A1 skrives en startværdi for μ eksempelvis 90. ▶ I celle B1 skrives =NORMFORDELING(90;A1;0,5/KVROD(12);1) ▶ Funktioner ▶ "Målsøgning" I "Angiv celle" skrives B1. I "Til Værdi" skrives 0,025. I "Ved ændring af celle" skrives A1. Resultat 90,2841

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

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

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

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.

Er spredningen eksakt kendt gælder der generelt (udtrykt ved α) formlen

$$\bar{x} - u_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + u_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \qquad (2)$$

Eksempel 3.3. Konfidensinterval hvis spredningen er kendt eksakt

Lad os antage, at vi spredningen for en population kendes eksakt til $\sigma = 5{,}8$

1) Bestem et 95% konfidensinterval for en stikprøve på 5 elementer, der har gennemsnittet $\bar{x} = 7.74$
2) Bestem et 95% konfidensinterval for en stikprøve på 30 elementer, der har gennemsnittet $\bar{x} = 12.65$

Løsning:

"Radius" r i et 95% konfidensinterval er $r = u_{0.975} \cdot \frac{\sigma}{\sqrt{n}} = 1.96 \cdot \frac{5.8}{\sqrt{n}}$

$u_{0.975}$ kan beregnes af Excel ved $u_{0.975}$ = NORMINV(0,975;0;1) = 1,959961,eller slås op i tabel 1

1) $r = 1.96 \cdot \frac{5.8}{\sqrt{5}} = 5.08$.

Lettere er det at finde radius r ved

På værktøjslinien foroven: Tryk på = eller f_x ▶ Vælg kategorien "Statistisk" ▶ Vælg "konfidensinterval" ▶ udfylde menuen : KONFIDENSINTERVAL(0,05;5,8;5)=5,08

95% konfidensinterval: $7.74 - 5.08 \le \mu \le 7.74 + 5.08 \Leftrightarrow \underline{\underline{2.66 \le \mu \le 12.82}}$

2) $r = 1.96 \cdot \frac{5.8}{\sqrt{30}} = 2.08$

95% konfidensinterval: $12.65 - 2.08 \le \mu \le 12.65 + 2.08 \Leftrightarrow \underline{\underline{10.57 \le \mu \le 14.73}}$

Vi ved derfor med 95% sikkerhed, at populationens sande middelværdi ligger indenfor disse intervaller[2]. ◆

[2] Mere præcist, at af de 100 stikprøver med tilhørende 95% konfidensintervaller, vil i middel kun 5 af disse intervaller ikke indeholde den sande værdi.

3.2.2. 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.
Hvis stikprøvestørrelsen er over 30 er denne usikkerhed dog uden væsentlig betydning, så i sådanne tilfælde kan man i formel (1) og (2) blot erstatte σ med s.

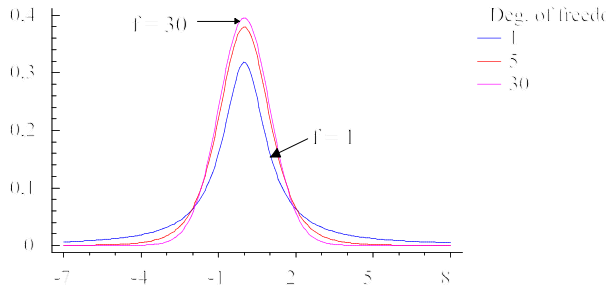
Er stikprøvestørrelsen under 30 bliver denne usikkerhed på s så stor, at man i formel (1) må erstatte U- fraktilen $u_{0.975}$ med en såkaldt t - fraktil $t_{0.975,f}$.

eller udtrykt ved α i formel (2) erstatte U- fraktilen $u_{1-\frac{\alpha}{2}}$ med t - fraktilen $t_{1-\frac{\alpha}{2},f}$.

t-fordelinger

En t - fordeling har samme klokkeformede udseende som en U - fordeling (en normalfordeling med middelværdi 0 og spredning 1)
I modsætning til U - 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 meget lille.
Er f lille bliver t - fordelingen bredere end U - fordelingen.

Grafen nedenfor viser tæthedsfunktionen for t-fordelingerne for f = 1, 5 og 30.



Som det ses af formlen for konfidensinterval har vi kun brug for at beregne t - fraktiler.

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

Eksempel 3.4. Beregning af t-fraktiler.

Find fraktilerne $t_{0.975,12}$ og $t_{0.025,12}$.

Løsning:

Af symmetrigrunde (se figuren) er de 2 fraktiler lige store med modsat fortegn, dvs. $t_{0.025,12}$ =- $t_{0.975,12}$

Excel: På værktøjslinien foroven: Tryk på = eller f_x ▶ Vælg kategorien "Statistisk" ▶ Vælg "TINV"

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

Bemærk: TINV(α ;*f*) udregner den fraktil, der svarer til 1 - $\frac{\alpha}{2}$

Sætter vi således $\alpha = 5\%$ fås $t_{0.975}$

$t_{0.025,12}$ = TINV(0.05;12) = 2,178813

$t_{0.025,12}$ = - 2,178813

Er spredningen ukendt er et 95 % konfidensinterval bestemt ved formlen:

$$\bar{x} - t_{1-\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{1-\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}} \qquad (3)$$

Er spredningen ukendt er formlen for et konfidensinterval (udtrykt ved α)

$$\bar{x} - t_{1-\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{1-\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}} \qquad (4)$$

Eksempel 3.5. Konfidensinterval, hvis spredningen ikke er kendt eksakt.

En forstmand er interesseret i at bestemme middelværdien af diameteren af voksne egetræer i en bestemt fredet skov.

Der blev målt diameteren på 7 tilfældigt udvalgte egetræer (i 1 meters højde over jorden)

Resultatet ses i følgende skema.

diameter (cm)	64.0	33.4	45.8	56.0	51.5	29.2	63.7

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

Løsning:

Data indtastes i Excel i cellerne A1 til A7

1) På værktøjslinien foroven: Tryk på f_x ▶ Vælg kategorien "Statistisk" ▶ Vælg "middel"

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

   $\bar{x}$ =MIDDEL(A1:A7)= 49,08571

   Tilsvarende fås s = =STDAFV(A1:A7)=13,7957

2) $\bar{x} \pm t_{1-\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}} = 49.086 \pm t_{0.975,7-1} \cdot \frac{13.796}{\sqrt{7}}$

Idet $\bar{x}$ er gemt i A8 og s i A9 fås
Nedre grænse = A8-TINV(0,05;6)*A9/KVROD(7) = 36,32681
Øvre grænse =A8+TINV(0,05;6)*A9/KVROD(7) = 61,84462
95% konfidensinterval: [36.33 ; 61.85]
Radius r i konfidensintervallet kan også findes ved i menuen i "Beskrivende Statistik" at afmærke "konfidensniveau for middelværdi"

	Kolonne1
Konfidensniveau(95,0%)	12,7589

95% konfidensinterval: [49.08-12.76 ; 49.08 + 12.76] =[36.32 ; 61.84]

3.2.3. Dimensionering

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

$r = u_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$. Løses denne ligning med hensyn til n fås

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

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.
Endvidere vil man som en første tilnærmelse antage, at antallet af gentagelser er over 30, så man kan bruge u-fordelingen. Det følgende eksempel illustrerer fremgangsmåden.

Eksempel 3.6. Dimensionering.
Forstmanden i eksempel 3.4 fandt, at konfidensintervallet der blev beregnet på basis af 7 træer var for bredt.
Han ønskede, at 95% konfidensintervallet højst skulle have en radius på 5 cm.
På basis af resultaterne i eksempel 3.3 sættes $s \approx 14$. Da samtidig $u_{0.975} \approx 2$ fås

$$n = \left(\frac{u_{0.975} \cdot s}{r} \right)^2 = \left(\frac{2 \cdot 14}{5} \right)^2 \approx 32$$

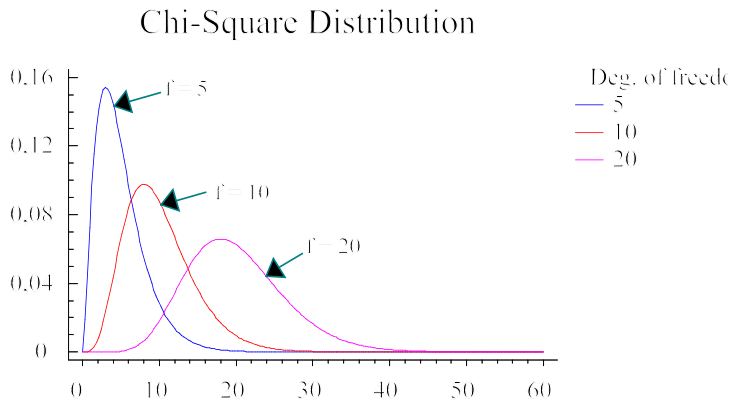
Da $n > 30$ er det rimeligt, at benytte en U- fordeling frem for en t-fordeling.
Der skal altså tilfældigt udvælges ca. 32 egetræer.
Da overslaget jo er afhængigt af om vurderingen af s er korrekt, bør man for en sikkerheds skyld vælge s lidt rigelig stor.

3.3. Konfidensinterval for spredning

Man kan vise, at et konfidensinterval for spredning er bestemt ved formlen $\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}}$, forudsat,

at middelværdien ikke er kendt eksakt.

Her er nævnerne fraktiler i den såkaldte χ^2 - fordeling (se figuren)



Fraktilerne kan beregnes i Excel

Eksempel 3.7. 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 middelværdi og spredning.

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

Løsning:

Excel:

Data indtastes i Excel i cellerne A1 til A12

1) På værktøjslinien foroven: Tryk på = eller f_x ▶ Vælg kategorien "Statistisk" ▶ Vælg "middel"

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

$\bar{x}$ =MIDDEL(A1:A12)= 6,3

Tilsvarende fås s = =STDAFV(A1:A7)= 0,316228

2) $\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}} \Leftrightarrow \frac{(12-1)\cdot 0.316228}{\chi^2_{0.975,11}} \le \sigma^2 \le \frac{(12-1)\cdot 0.316228}{\chi^2_{0.025,11}}$

Idet $\bar{x}$ er gemt i A13 og s i A14 fås

Nedre grænse = (12-1)*A14^2/CHIINV(0,025;11) = 0,050182

Øvre grænse =(12-1)*A14^2/CHIINV(0,975;11) = 0,288279

95% konfidensinterval for variansen: [0.0502 ; 0.288]

95% konfidensinterval for spredningen: $\sqrt{0.0502} \le \sigma \le \sqrt{0.2880} \Leftrightarrow 0.2241 \le \sigma \le 0.5366$

Bemærk: Excel beregner den "øvre hale". ◆

Opgaver

Opgave 3.1

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) Man fandt, at radius i konfidensintervallet var for stor.
   Bestem med tilnærmelse antallet af målinger der skal udføres, hvis radius højst skal være 15.

Opgave 3.2

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

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

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

Opgave 3.3

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

pH	8.54	7.89	8.50	8.21	8.15	8.32

Den benyttede pH-målemetode antages på baggrund af tidligere lignende målinger at give normalfordelte resultater.

1) Angiv et estimat for opløsningens middelværdi og spredning.
2) Angiv et 95% konfidensinterval for pH.
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 3.4

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

4.21	4.33	4.26	4.27	4.19	4.30	4.24	4.24	4.28	4.24

Angiv et 95%-konfidensintervaller for middelværdien af papirets vægt.

Opgave 3.5

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

108	102	107	98

Opstil et 95% konfidensinterval for middelværdien af personens alkoholkoncentration.

Opgave 3.6 = opgave 3.1 fortsat.
Find et 95% konfidensinterval for trykstyrkens spredning.

Opgave 3.7 = opgave 3.2 fortsat.
Find ud fra stikprøven et 99% konfidensinterval for diameterens spredning.

Opgave 3.8 = opgave 3.4 fortsat.
Find et 95% konfidensinterval for spredningen af papirets vægt.

Opgave 3.9 = opgave 3.5 fortsat.
Opstil et 95% konfidensinterval for spredningen af personens alkoholkoncentration.

4. Sandsynlighedsregning

4.1 Indledning

Vi har i det foregående i forbindelse med indføringen af normalfordelingens tæthedsfunktion "defineret" sandsynligheden $P(a \le X \le b)$. For at kunne behandle andre statistiske fordelings-funktioner er det nødvendigt at kende visse grundlæggende regneregler for sandsynlighed.

4.2. Sandsynlighed

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

Eksempler på tilfældige eksperimenter

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

2) Består eksperimentet i, at vi fra et skib affyrer et skud mod et mål, ved vi, at at enten rammer vi målet, eller også gør vi det ikke (Udfaldsrummet U = [ramme ; ikke ramme]), men vi kan ikke forudsige resultatet.

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

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

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

Eksempel 4.1. De relative hyppigheders stabilitet
Et eksperiment består i at kaste en terning, og hændelsen A består i at få et lige øjental. Terningen kastes nu 100 gange, og man får et lige øjental 55 gange. Eksperimentet udføres igen 100 gange, og man får A 47 gange. Igen kastes 100 gange, og man får nu A 57 gange. Til sidst kastes 100 gange, og man får A 40 gange.
Eksperimentet foretages nu i serier på 1000 gange, hvor man hver gang optæller antal gange A forekommer. Resultaterne vises i følgende tabel:

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

Det ses, at med 1000 gentagelser er de relative hyppigheder tættere samlet (ligger mellem 48,6% og 50,9%) end hvis man kun kastede 100 gange (mellem 40% og 55%). Hvis terningen var en ægte terning(fuldstændig homogen og symmetrisk), måtte man på forhånd forvente, at det tal, som de relative hyppigheder grupperede sig omkring, var tallet 0.5. Man vil derfor sige, at sandsynligheden for at få et lige øjental er 0.5, eller kort $P(A) = 0.5$. ◆

4.3 Regneregler for sandsynligheder

I dette afsnit vil følgende eksempel blive benyttet til illustration af definitioner og begreber.

Eksempel 4.2. Gennemgående eksempel.

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

◆

Komplemtærmængden til A benævnes $\overline{A}$ og er mængden af alle udfald i udfaldsrummet U, der ikke er i A (den skraverede mængde på figur 4.1).

Eksempelvis er $\overline{A}$ i eksempel 4.2 mængden af alle glasfigurer, der ikke er misfarvet.
Idet $P(A) = 0.1$ ses umiddelbart, at $P(\overline{A}) = 1 - P(A) = 0.9$.

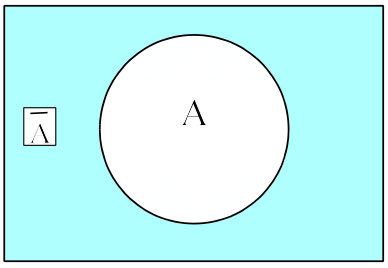


Fig. 4.1. *Komplementærmængde*

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

Fællesmængden til A og B benævnes $A \cap B$ og er mængden af alle udfald i udfaldsrummet U, der tilhører **både A og B** (Den skraverede mængde i figur 4.2).

Eksempelvis er $A \cap B$ i eksempel 4.2 mængden af alle glasfigurer, der både er misfarvede og ridsede.
Af eksemplet følger, at $P(A \cap B) = 0.01$.

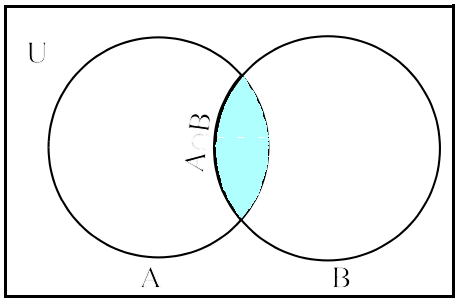


Fig 4.2. *Fællesmængde*

Foreningsmængden af A og B benævnes $A \cup B$ og er mængden af alle udfald i udfaldsrummet U, der **enten tilhører A eller B eventuelt dem begge** (den skraverede mængde på figur 4.3)

Eksempelvis er $A \cup B$ i eksempel 4.2 mængden af alle glasfigurer, der enten er misfarvede eller ridsede eventuelt begge dele.

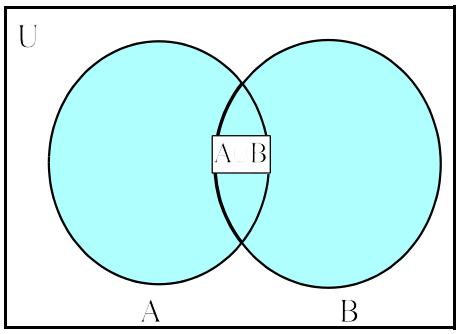


Fig. 4.3 *Foreningsmængde*

Ved betragtning af de nedenfor anførte grafiske afbildninger (arealbetragtning) ses umiddelbart, at der gælder følgende

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

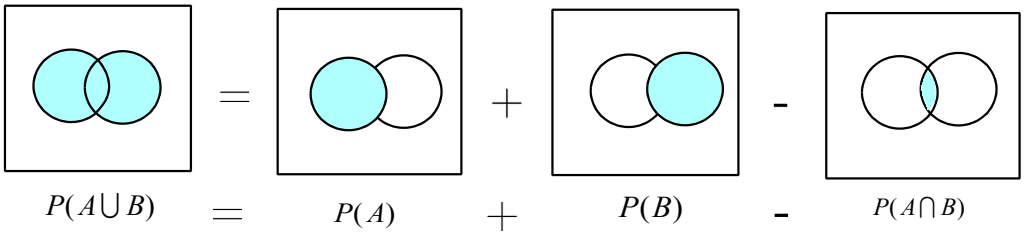


Fig.4.4 *Additionssætning*

I eksempel 4.2 er $P(A \cup B) = 0.1 + 0.2 - 0.01 = 0.29.$

4.4. Betinget sandsynlighed

For nogle hændelser A og B gælder, at $P(A \cap B) = P(A) \cdot P(B)$, men denne formel gælder ikke generelt. Eksempelvis er i eksempel 4.2 $P(A) \cdot P(B) = 0.1 \cdot 0.2 = 0.02 \neq P(A \cap B)$.

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

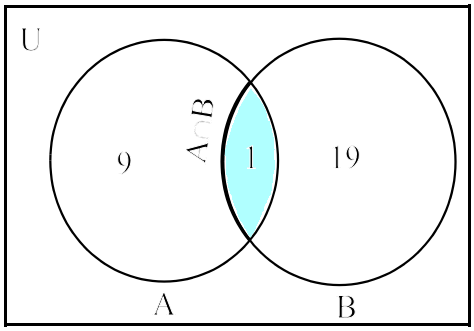


Fig. 4.4 *Taleksempel*

For at forklare den følgende definition, vil vi simplificere eksempel 4.2, idet vi antager, at den daglige produktion er 100 glasfigurer. I så fald er der 10 misfarvede figurer, 20 ridsede figurer, og 1 figur der er både misfarvet og ridset.

Hvis vi begrænser vort udfaldsrum til A, så er $P(B\,|\,A) = \frac{1}{10} = \frac{\frac{1}{100}}{\frac{10}{100}} = \frac{P(A \cap B)}{P(A)}$.

Denne beregning begrunder rimeligheden i følgende definition:

Den af A **betingede sandsynlighed** for B $P(B\,|\,A)$ (eller 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 4.2 fås $P(A \cap B) = P(A) \cdot P(B\,|\,A) = 0.1 \cdot 0.1 = 0.01$.

Eksempel 4.3: Betinget sandsynlighed.
En beholder indeholder 3 røde og 3 hvide kugler. Vi udtrækker successivt 2 kugler fra urnen.
Vi betragter følgende 2 hændelser:
A: Den først udtrukne kugle er rød. B: Den anden udtrukne kugle er rød.
Beregn $P(A \cap B)$ hvis
1) kugleudtrækningen foregår, ved at den først udtrukne kugle lægges tilbage før den anden udtrækkes.
2) kugleudtrækningen foregår, ved at den først udtrukne kugle **ikke** lægges tilbage før den anden udtrækkes.

Løsning
1) Her er $P(B\,|\,A) = \frac{3}{6}$ 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: $P(B|A) = \dfrac{P(B) \cdot P(A|B)}{P(A)}$

Bevis:

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

♦

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

Eksempel 4.4 (Bayes sætning)

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

Ved lodtrækning vælges en elev.

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

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

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

Løsning:

Vi definerer følgende hændelser:

T: Den udtrukne er tekniker

K: Den udtrukne er en kvinde.

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

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

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

Vi har så følgende skema

	Kvinder	I alt
Operativ	12	60
Teknisk	4	40

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

♦

4.5. Statistisk uafhængighed.

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

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

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

Opgaver

Opgave 4.1

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.

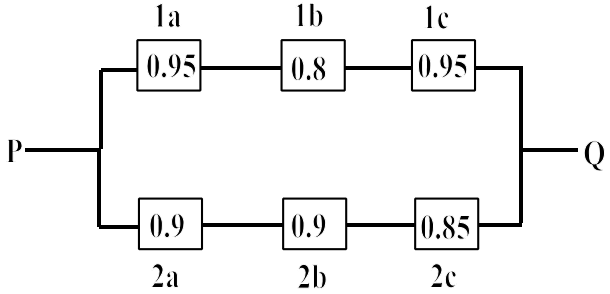
C: Husstanden holder begge aviser .
D: Husstanden holder kun den lokale avis.
E: Husstanden holder mindst én af aviserne.
F: Husstanden holder ingen avis
G: Husstanden holder netop én avis.

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

H: Husstanden holder den lokale avis.

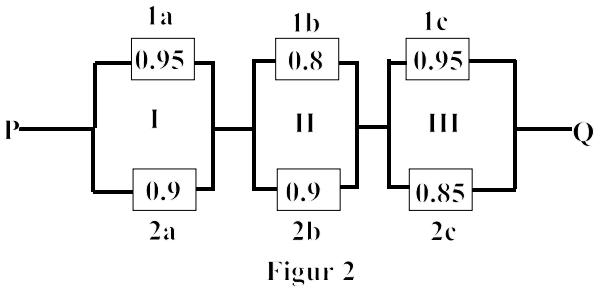
Opgave 4.2

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 4.3
Tre skytter skyder hver ét skud mod en skydeskive. De har træffesandsynligheder 0.75, 0.50 og 0.30.
Beregn sandsynligheden for
1) ingen træffere, 2) én træffer, 3) to træffere, 4) tre træffere.

Opgave 4.4
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 4.5
En virksomhed fremstiller en bestemt slags apparater. Hvert apparat er sammensat af 5 komponenter. Heraf er 3 tilfældigt udvalgt blandt komponenter af typen a og 2 blandt komponenter af typen b. Det vides, at 10% af a-komponenterne er defekte og 20% af b-komponenterne er defekte. Et apparat fungerer hvis og kun hvis det ikke indeholder nogen defekt komponent.
Der udtages på tilfældig måde et apparat fra produktionen. Lad os betragte hændelserne:
A: Det udtagne apparat indeholder mindst 1 defekt a-komponent.
B: Det udtagne apparat indeholder mindst 1 defekt b-komponent.
1) Find $P(A), P(B)$ og $P(A \cap B)$.
2) Find sandsynligheden for, at et apparat, der på tilfældig måde udtages af produktionen ikke fungerer.
3) Et apparat udtages på tilfældig måde fra produktionen og det konstateres ved afprøvning at det ikke fungerer. Find sandsynligheden for, at apparatet ikke indeholder nogen defekt a-komponent.

Opgave 4.6
To skytter konkurrerer ved en turnering. De har hver én patron og skyder mod en skive som giver 10 point, hvis et centralt område af skiven rammes og ellers 5 point. Rammes skiven ikke noteres 0 point.
Skytte A's dygtighed kan beskrives ved, at han i et skud har samme sandsynlighed for at få 10 points, 5 points eller 0 points.
Skytte B er dygtigere, idet hans sandsynligheder for at ramme er givet ved

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

B har imidlertid fået en defekt patron med, der har sandsynligheden 50% for at fungere.
1) Idet X betegner det af A opnåede antal points og Y det af B opnåede antal points, ønskes tæthedsfunktionen for X og Y beregnet.
2) Find $E(X), E(Y), \sigma(X)$ og $\sigma(Y)$.
3) Beregn sandsynligheden for, at A vinder.
4) Det oplyses, at A vandt konkurrencen. Beregn sandsynligheden for, at B opnåede 5 points.

5. Kombinatorik

5.1. 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**.

5.2. 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 5.1. Multiplikationsprincippet

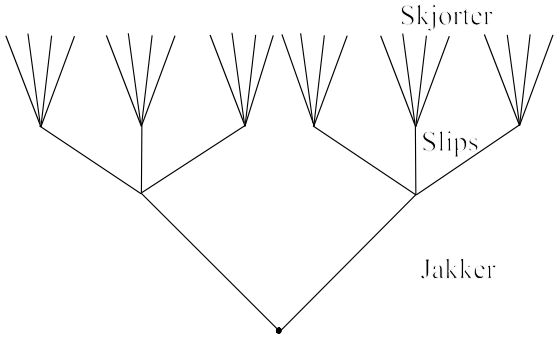
En mand ejer 2 forskellige jakker, 3 slips og 4 forskellige fabrikater skjorter.
På hvor mange forskellige måder kan han sammensætte sin påklædning af jakke, slips og skjorte.

Løsning:

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

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

Man kunne illustrere løsningen ved følgende "forgreningsgraf"



♦

5.3 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

5.3.1 Uden tilbagelægning

Eksempel 5.2. Ordnet uden tilbagelægning

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

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

Løsning:

1) En formand vælges blandt 10 kandidater — 10 valgmuligheder
   En Kasserer vælges blandt de resterende 9 kandidater — 9 valgmuligheder
   Da der for hvert valg af formand er 9 muligheder for kasserer, følger af multiplikationsprincippet, at det totale antal forskellige bestyrelser er $10 \cdot 9 = \underline{\underline{90}}$.
2) Analogt fås ifølge multiplikationsprincippet at antal forskellige bestyrelser er
   $10 \cdot 9 \cdot 8 \cdot 7 = \underline{\underline{5040}}$
   Excel: På værktøjslinien foroven: Tryk på f_x ▶ Vælg kategorien "Statistisk" ▶ Vælg "Permut" ▶ udfylde menuen . Resultat: =PERMUT(10;4) = 5040 ♦

Eksempel 5.2 begrunder følgende definition

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

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

♦

5.3.2 Med tilbagelægning

Eksempel 5.3. Ordnet, med tilbagelægning

I en forening skal 4 tillidshverv fordeles mellem 10 personer. En person kan godt have flere tillidshverv. På hvor mange forskellige måder kan disse hverv fordeles.?

Løsning:

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

♦

5.4. Uordnet stikprøveudtagelse

Eksempel 5.4 Uordnet uden tilbagelægning

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

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

Hvor mange forskellige stikprøver kan forekomme?

Løsning:

Antallet er ikke flere end man kan foretage en simpel optælling:

$\{k_1,k_2,k_3\},\{k_1,k_2,k_4\}\{k_1,k_2,k_5\}\{k_1,k_3,k_4\}\{k_1,k_3,k_5\}\{k_2,k_3,k_4\}\{k_2,k_3,k_5\}\{k_2,k_4,k_5\}\{k3,k_4,k_5\}$

Antal stikprøver = 10

♦

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

Definition af kombination

Lad M være en mængde med *n* elementer.

En delmængde af M med *r* elementer kaldes en **kombination** af med *r* elementer fra M.

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

Sætning 5.1 (Antal kombinationer).

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

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

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

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

Vi har følgelig, at $P(5,3) = K(5,3) \cdot 3! \Leftrightarrow K(5,3) = \frac{P(5,3)}{3!} \Leftrightarrow K(5,3) = \frac{5 \cdot 4 \cdot 3}{3!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{3! \cdot 2!} = \frac{5!}{3! \cdot 2!}$ ◆

Eksempel 5.5. Antal kombinationer

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

Løsning:

Antal måder man kan sammensætte bestyrelsen er

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

Excel: På værktøjslinien foroven: Tryk på = eller f_x ▶ Vælg kategorien "Statistisk" ▶ Vælg "Kombin" ▶ udfylde menuen . Resultat: ==KOMBIN(10;4) = 210 ◆

5.5 Hypergeometrisk fordeling

Af særlig interesse er den såkaldte "hypergeometriske fordeling", som bl.a. finder anvendelse ved kvalitetskontrol af varepartier (jævnfør eksempel 5.9),ved markedsundersøgelser, hvor man uden tilbagelægning udtager en repræsentativ stikprøve på eksempelvis 500 personer

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

Eksempel 5.6. Hypergeometrisk fordeling

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

Løsning:

X = antal kvinder i bestyrelsen
At der skal være netop 1 kvinde i bestyrelsen forudsætter, at vi udtager 1 kvinde ud af de 5 kvinder og 3 mænd ud af de 8 mænd.
At udtage 1 kvinde ud af 5 kvinder kan gøres på K(5,1) måder
At udtage 3 mænd ud af 8 mænd kan gøres på K(8,3) måder.
Antal gunstige udfald er ifølge multiplikationsprincippet $K(5,1) \cdot K(8,3)$
Det totale antal udfald fås ved at udtage 4 personer ud af de 13 kandidater
Dette kan gøres på K(13,4) måder.

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

Definition af hypergeometrisk fordeling.
Lad der i en beholder befinde sig N kugler, hvoraf M er defekte.
En kugle udtrækkes og undersøges.
Dette gentages *n* gange **uden mellemliggende tilbagelægning**
Lad X være antallet af defekte kugler, som udtrækkes. Der gælder da

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

Eksempel 5.6 (fortsat)
I eksempel 5.6 benyttede vi den hypergeometriske fordeling for N = 13, M = 5 , n = 4 og $x = 1$.
I Excel kan beregningen foretages ved
På værktøjslinien foroven: Tryk på f_x ▶ Vælg kategorien "Statistisk" ▶ Vælg "Hypergeo" ▶ udfylde menuen
. Resultat: =HYPGEOFORDELING(1;5;4;13) = 0,391608 = 39.16%

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

Eksempel 5.7. 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:
Lad *X* være antallet af defekte blandt de 25 komponenter
Vi har: $p = P(X=0) + P(X=1)$.

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

Vi har altså $p = 0.6512 + 0.2876 = 0.9388 = \underline{93.88\%}$.

Opgaver

Opgave 5.1.

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

Opgave 5.2.

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

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

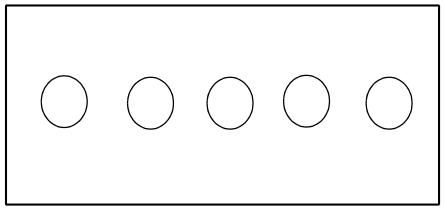
Opgave 5.3.

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

Opgave 5.4.

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

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



Opgave 5.5

Normale personbilers indregistreringsnumre består af to bogstaver og et nummer mellem 20000 og 59999 .

Lad os antage, at man er nået til numre der begynder med UV. Et eksempel på en nummerplade er da UV 54755

Hvad er sandsynligheden for, at en nyindregistreret bil får et registreringsnummer med lutter forskellige cifre, når vi antager, at alle cifre har samme sandsynlighed?

Opgave 5.6

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 5.7

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 5.8.
Bestem antallet af 5-cifrede tal, der kan skrives med to l-taller, et 2- tal og to 3-taller.

Opgave 5.9
Hvor mange forskellige telefonnumre på 8 cifre kan man danne, når første ciffer ikke må være nul?

Opgave 5.10
En beholder indeholder 3 hvide, 6 røde og 3 sorte kugler
3 kugler udtrækkes tilfældigt uden tilbagelægning.
Find sandsynligheden for at de er af samme farve.

Opgave 5.11
Fra et sædvanligt spil kort udtrækkes på tilfældig måde 3 kort uden tilbagelægning. Bestem sandsynlighederne for hver af hændelserne
A: Der udtrækkes kun 8'ere.
B: Der udtrækkes lutter hjerter.
C: Der udtrækkes 2 sorte og 1 rødt kort.

Opgave 5.12
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 5.13
Ved en lodtrækning fordeles 3 gevinster blandt 25 lodsedler. En spiller har købt 5 lodsedler. Beregn sandsynligheden for hver af følgende hændelser:
1) Spilleren vinder alle tre gevinster.
2) Spilleren vinder ingen gevinster.
3) Spilleren vinder netop én gevinst.

Opgave 5.14
I en urne findes 2 blå, 3 røde og 5 hvide kugler. 3 gange efter hinanden optages tilfældigt en kugle fra urnen uden mellemliggende tilbagelægning.
1) Find sandsynligheden for hændelsen A, at der højst optages 2 hvide kugler,
2) Find sandsynligheden for hændelsen B, at de optagne kugler har hver sin farve.
3) Find sandsynligheden for, at de tre kugler har samme farve,

Opgave 5.15
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.

6 BINOMIALFORDELING

6.1. Indledning

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

6.2. Definition og beregning

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

Eksempel 6.1. En binomialfordelt variabel.
En skytte har 15% sandsynlighed for at ramme målet.
Skytten skyder 6 gange. Hvad er sandsynligheden for at skytten har netop 2 træffere.
Lad X være antallet af træffere blandt de 6 skud
Vi ønsker at finde sandsynligheden for at finde netop 2 træffere blandt disse 6, det vil sige $P(X=2)$.
Løsning:
Lad et eksperiment være at skyde et skud.
Resultatet af eksperimentet har to udfald: træffer , forbier.
Eksperimentet gentages 6 gange uafhængigt af hinanden.
Der er en bestemt sandsynlighed for at få en træffer, nemlig $p = 0.15$.
Lad t være det udfald at få en træfferdefekt, og f være det udfald at få en forbier.
Et af de ønskede forløb med 2 træffere vil eksempelvis være *t, f, t, f, f, f, f.*
Dette forløb må have sandsynligheden
$0.15\cdot(1-0.15)\cdot 0.15\cdot(1-0.15)\cdot(1-0.15)\cdot(1-0.15) = 0.15^2\cdot(1-0.15)^4$.
Et andet gunstigt forløb kunne være *f, f, t, f, t, f* med sandsynligheden
$(1-0.15)\cdot(1-0.15)\cdot 0.15\cdot(1-0.15)\cdot 0.15\cdot(1-0.15) = 0.15^2\cdot(1-0.15)^4$
Vi ser, at alle gunstige forløb har samme sandsynlighed.
Antal forløb må være lig antal måder man kan placere to *t*'er på 6 tomme pladser (eller antal måder man kan tage 2 kugler ud af en mængde på 6). Dette ved vi kan gøres på K(6,2) måder.
Vi får følgelig, at $p = K(6,2)\cdot 0.15^2\cdot(1-0.15)^4 = 0.1762 = \underline{\underline{17.62\%}}$

◆

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

DEFINITION af binomialfordeling.
1) Lad et tilfældigt eksperiment have 2 udfald "succes" og "fiasko"
2) Lad eksperimentet blive gentaget n gange uafhængigt af hinanden, og lad sandsynligheden for succes være en konstant p
Lad X være antallet af succeser blandt de n gentagelser
Der gælder da: $P(X=x) = K(n,x)\cdot p^x\cdot(1-p)^{n-x}$ for $x\in\{0,1,2,\ldots,n\}$
X siges at være binomialfordelt b (n, p).

Eksempel 6.2 (beregning af binomialfordeling)
I eksempel 6.1 fandt vi, at X var binomialfordelt $b(6, 0.01)$.
$P(X = 1)$ beregnes i Excel på følgende måde:
På værktøjslinien foroven: Tryk på f_x ▶ Vælg kategorien "Statistisk" ▶ Vælg "Binomialfordeling" ▶ udfylde menuen . Resultat: ==BINOMIALFORDELING(2;6;0,01;FALSK) = 0,0014410 = 0.14%

◆

Approksimation af hypergeometrisk fordeling med binomialfordeling.
Den hypergeometriske fordeling anvendes sædvanligvis ved kvalitetskontrol, da man udtager stikprøven uden tilbagelægning. Hvis man i stedet efter at have taget et emne op og undersøgt det lagde emnet tilbage, så var der jo en fast sandsynlighed for at få en defekt. I et sådant tilfælde var fordelingen derfor binomialfordelt. Hvis man udtager en lille stikprøve af størrelsen *n af* en stor mængde af størrelsen N, vil sandsynligheden for at få en defekt ikke ændre sig meget hvad enten man lægger tilbage eller ej. For de fleste anvendelser kan man derfor med en passende nøjagtighed erstatte den hypergeometriske fordeling med binomialfordelingen,, hvis stikprøvestørrelsen n er mindre end eller lig 10% af partistørrelsen N ($\frac{n}{N} \le \frac{1}{10}$).

Eksempel 6.3. Hypergeometrisk fordeling approksimeret med binomialfordeling .
I eksempel 5.8 betragtede vi følgende situation.
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.
Løsning:
Lad X være antallet af defekte blandt de 25 komponenter
Vi har: $p = P(X \le 1)$
Da $\frac{n}{N} = \frac{25}{600} < \frac{1}{10}$ kan approksimeres med binomialfordelingen $b\left(25, \frac{10}{600}\right)$.
$P(X \le 1)$ = BINOMIALFORDELING(1,25,1 / 60,1) = 0.9353
Benyttede vi den hypergeometriske fordeling fandt vi 93.88%. Denne forskel på 0.35% har næppe praktisk betydning.

◆

Middelværdi og spredning for binomialfordeling *b(n,p)*
Binomialfordelingen har middelværdien $\mu = n \cdot p$ *og spredningen* $\sigma = \sqrt{n \cdot p \cdot (1-p)}$.
Heraf fås (ved division med n), at p har spredningen $\sigma(p) = \sqrt{\frac{p \cdot (1-p)}{n}}$.

Et bevis vil ikke blive foretaget her.

Eksempel 6.4: Sandsynlighedsfunktion for binomialfordeling .

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

En repræsentativ stikprøve på 15 husstande udtages.

Lad X = antal husstande med parabol ud af 15

Da man må antage, at man ikke spørger den samme husstand to gange er X fordelt hypergeometrisk med $N = 1000$, $M = 700$ og $n = 15$.

Da $\frac{n}{N} = \frac{15}{1000} < 0.1$ kan man tillade sig at approksimere med binomialfordelingen b(15, 0.70).

(antallet N af indbyggere i kommunen er så stort, at sandsynligheden ikke ændrer sig, fordi man har udtaget op til 15 husstande).

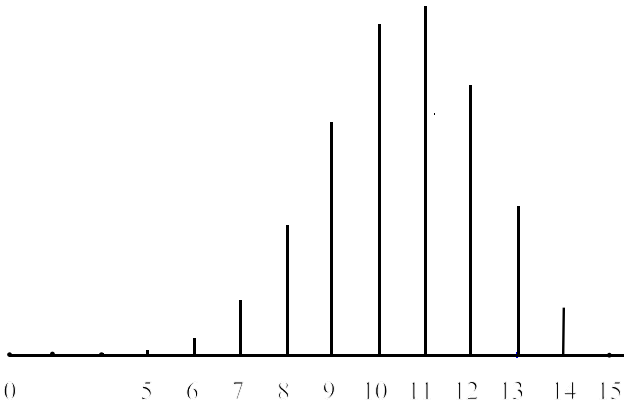
1) Tegn sandsynlighedsfunktionen for X (idet X antages binomialfordelt)
2) Beregn middelværdi og spredning for X

Løsning:

1) Da X er en "diskret" variabel, der kun antager hele værdier tegnes et stolpediagram.
   Vi beregner værdierne ved at benytte Excel, eksempelvis
   $P(X = 9)$ = BINOMIALFORDELING(9;15;0,7;FALSK) = 0,147
   Vi får følgende tabel:

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

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



Stolpediagram for binomialfordelingen

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

6.3. Konfidensinterval for p .

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

Eksempel 6.5. Opinionsundersøgelse.
Ved valget i 2004 stemte 25.9% af vælgerne på socialdemokraterne.
I en opinionsundersøgelse svarede 1035 vælgere på spørgsmålet om hvilket parti det var mest sandsynligt de ville stemme på hvis der var valg i morgen.

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

Løsning:

a) Idet 25.9% af 1035 er ca. 268, og 24.8% af 1035 er ca. 257.
Lad X = antal vælgere der svarer, at de vil stemme på socialdemokraterne ud af 1035 vælgere.
X antages binomialfordelt med n = 1035 og p ukendt.
Under forudsætning af at partiet har samme tilslutning som ved valget vil vi beregne sandsynligheden for at man ved opinionsundersøgelsen får 257 stemmer eller færre.
$P(X \le 257)$ ==BINOMIALFORDELING(257;1035;0,259;1) = 0.2275 = 22.75%
Hvis socialdemokraterne har samme tilslutning som ved valget er der altså ca. 23% sandsynlighed for at man ved en opinionsundersøgelse ville få samme resultat, dvs. at 257 (eller færre) ville sige, de ville stemme på partiet. Man kan derfor ikke med rimelig fastslå, at denne opinionsundersøgelse viser, at partiet er gået tilbage i tilslutning i forholdet til valget.
b) Hvis socialdemokraterne ved opinionsundersøgelsen kun havde fået 23% af stemmerne , svarende til ca. 238 stemmer, så finder vi tilsvarende, at
$P(X \le 238)$ = BINOMIALFORDELING(238;1035;0,259;1) = 0.017 = 1.7%
Nu er der kun 1.7% sandsynlighed for at man ville få et sådant resultat, hvis tilslutningen var uændret, så umiddelbart må man konkludere, at tilslutningen er faldet.

◆

Som det fremgår af eksempel 6.5 er spørgsmålet om at tilslutningen er faldet, afhængig af hvor sikker man vil være.
Man vil sædvanligvis ønske at angive et "usikkerhedsinterval" som angiver, at den "sande" tilslutning p til partiet med 95% sikkerhed ligger indenfor dette interval. Vil man være mere sikker, så kan man jo i stedet vælge 99% eller 99.9%, men da regningerne i princippet er de samme, vil vi i det følgende vælge 95%.

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

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

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

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

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

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

Når man skal lave et konfidensinterval benytter man sig af, at for store stikprøvestørrelser, vil et estimat $\hat{p}$ for parameteren p være tilnærmelsesvis normalfordelt.

Dette begrunder følgende formel for konfidensinterval:

95% konfidensinterval for *p* i binomialfordelt variabel .

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

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

$$\hat{p} - u_{0.975} \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}} \leq p \leq \hat{p} + u_{0.975} \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}}$$

Eksempel 6.6 Beregning af konfidensinterval

I eksempel 6.5 svarede 1035 vælgere på spørgsmålet om hvilket parti det var mest sandsynligt de ville stemme på hvis der var valg i morgen. 24.8% svarede, at de ville stemme på Socialdemokraterne.

Opstil et 95% konfidensinterval for sandsynligheden p for at man vil stemme på socialdemokraterne.

Løsning:

Da $\hat{p}$ =0.248 fås $n \cdot \hat{p} \cdot (1-\hat{p}) = 1035 \cdot 0.248 \cdot (1-0.248) \approx 193 \geq 10$

Vi har nu : $0.248 - u_{0.975} \cdot \sqrt{\frac{0.248 \cdot (1-0.248)}{1035}} \leq p \leq 0.248 - u_{0.975} \cdot \sqrt{\frac{0.248 \cdot (1-0.248)}{1035}}$

Indsættes $u_{0.975} = 1.96$ fås $0.248 - 0.026 \leq p \leq 0.248 + 0.026 \Leftrightarrow 0.222 \leq p \leq 0.274$

Vi ser altså, at da valgresultatet lå på 25.9%, så kan man ikke med 95% sikkerhed sige, at man vil få et ringere resultat, hvis der var valg "i morgen".

Hvis betingelsen ikke er opfyldt (stikprøvestørrelsen n er for lille) kan man eventuelt benytte følgende (mere besværlige) metode.

Eksempel 6.7. Beregning af konfidensinterval hvis betingelserne ikke er opfyldt

I forbindelse med et reklamefremstød ønskede man at undersøge om borgerne i en mindre by havde set en bestemt reklame. Man spurgte et antal tilfældigt udvalgte husstande, og af 50 svar havde 10 set reklamen.

Opstil et 95% konfidensinterval for sandsynligheden p for at man har set reklamen.

Løsning:

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

Da $n \cdot \hat{p} \cdot (1-\hat{p}) = 50 \cdot 0.2 \cdot 0.8 = 8 < 10$ er betingelse 2 ikke opfyldt.

Vi finder nu den øvre grænse i konfidensintervallet ved at lade $\hat{p}$ stige indtil $P(X \le 10) \approx 0.025$

Excel: I celle A1 skrives en startværdi for p eksempelvis 0,3.
▶ I celle B1 skrives ==BINOMIALFORDELING(10;50;A1;SAND) ▶ Funktioner ▶ "Målsøgning"
I "Angiv celle" skrives B1. I "Til Værdi" skrives 0,025. I "Ved ændring af celle" skrives A1.
Resultat 0,336496

Dernæst findes nedre grænse ved at lade $\hat{p}$ falde, indtil $P(X \ge 10) = 1 - P(X \le 9) \approx 0.025$

I celle A1 skrives en startværdi for p eksempelvis 0,15.
▶ I celle B1 skrives ==1-BINOMIALFORDELING(9;50;A1;SAND) ▶ Funktioner ▶ "Målsøgning"
I "Angiv celle" skrives B1. I "Til Værdi" skrives 0,025. I "Ved ændring af celle" skrives A1.
Resultat 0,10048

Konfidensinterval: [0.100; 0.336]

6.4. Dimensionering

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 i konvergensintevallet er $r = u_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}}$.

Løses denne ligning med hensyn til n fås

$$n = \left(\frac{u_{1-\frac{\alpha}{2}}}{r} \right)^2 \hat{p} \cdot (1-\hat{p})$$

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 6.8. Dimensionering.

I den i eksempel 6.6 nævnte 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 $n = \left(\frac{u_{0.975}}{r}\right)^2 \hat{p} \cdot (1-\hat{p}) = \left(\frac{1.96}{0.02}\right)^2 \frac{1}{2} \cdot \frac{1}{2} = \underline{\underline{2401}}$

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

$n = \left(\frac{u_{0.975}}{r}\right)^2 \hat{p} \cdot (1-\hat{p}) = \left(\frac{1.96}{0.02}\right)^2 0.3 \cdot 0.7 = \underline{\underline{2017}}$

Opgaver

Opgave 6.1
Under en skydeøvelse viser det sig, at en premierløjtnant rammer et mål med 40% sandsynlighed. Premierløjtnanten affyrer 8 skud.
a) Find sandsynligheden for 3 træffere.
b) Find sandsynligheden for at få mindst 3 træffere.

Opgave 6.2
På flyvestationens hovedværksted har man fået oplyst, at sandsynligheden for en defekt bolt i en boltefabrikation er 0.1.
Man får en forsendelse med 400 bolte
a) Hvad er sandsynligheden for, at man i en stikprøve på 12 bolte finder mindst 1 defekt bolt.
b) Hvor mange defekte bolte vil der i middel være i forsendelsen.

Opgave 6.3
Under en øvelse affyrer en officer 80 skud mod et mål. På grund af meget vanskelige forhold er sandsynligheden for en træffer kun 0.05 i hvert forsøg.
Hvad er sandsynligheden for at officeren opnår mindst 5 træffere.

Opgave 6.4
Idet sandsynligheden for at ramme et større mål er 0.8, affyres 225 skud med en kanon. Målet anses for ødelagt, såfremt mindst 200 skud træffer det
Find sandsynligheden for at målet ødelægges.

Opgave 6.5
Når FLOS har fået anvist erhvervspraktikanter, har det desværre vist sig, at kun 60% af de anviste skoleelever dukker op. Forud for årets praktik har FLOS meddelt, at det kun er muligt at gennemføre praktiktjenesten, hvis der dukker mindst 12 praktikanter op.
Skoleelevernes "dukken op" er uafhængig af hinanden.
a) Hvad er sandsynligheden for at FLOS kan oprette et hold, hvis praktiktjenesten anviser 15 skoleelever til FLOS?
b) Hvor mange elever skal praktiktjenesten anvise, hvis der skal være 95% sandsynlighed for at kurset oprettes.

Opgave 6.6
Man udskifter i øjeblikket en ældre model fragmenteringsvest med en nyere. I lejrens depot udgør den ældre model 5% .
Ved en øvelse udleveres hurtigt og ganske tilfældigt 62 fragmenteringsveste til en deling. Hvad er sandsynligheden for, at flere end 5 fra delingen får udleveret en gammel model.

Opgave 6.7
En tipskupon har 13 kampe med 3 mulige tegn - 1, x og 2 - for hver kamp. En person bestemmer tegnet, der skal sættes for hver kamp, ved tilfældig udtrækning af en seddel fra 3 sedler med tegnene henholdsvis 1, x og 2. Angiv sandsynligheden for, at personen opnår netop 8 rigtige tippede kampe på sin kupon.

Opgave 6.8
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 6.9
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 6.10
I en urne er der et meget stort antal kugler, hvoraf de 70% er sorte. Fra urnen tages en stikprøve på 10 kugler. Find sandsynligheden for, at der i stikprøven er:
1) 10 sorte kugler
2) 6, 7 eller 8 sorte kugler
3) Mindst 7 sorte kugler

Opgave 6.11
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 6.12
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 6.13
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 6.14
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 6.15

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) 2,5% defekte møtrikker,
   c) 7,5% defekte møtrikker,
2) Bestem, for hvilken procentdel defekte møtrikker det ovennævnte parti (approksimativt) har 50% sandsynlighed for at blive godkendt af maskinfabrikken.

Opgave 6.16

Ved en fabrikation af plastikposer leveres disse i æsker med 100 poser i hver. Ved en godkendelseskontrol af et parti plastikposer udtages og undersøges en tilfældigt udtaget æske, og partiet godkendes, såfremt æsken højst indeholder én defekt pose.

Vi antager, at den løbende produktion af poser er således, at hver produktion med sandsynligheden 2% giver en pose, der er defekt.

Hvor stor er sandsynligheden for, at partiet under disse omstændigheder accepteres?

Opgave 6.17

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

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

Lad p være sandsynligheden for at få et defekt stykke legetøj.

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

Opgave 6.18

I rapporten "Analyse af elevkampagnen 2006" udarbejdet af "Forsvarets rekruttering" returnerede 604 personer et udsendt spørgeskema.

På side 10 er en opgørelse over hvilke medier der var udslagsgivende for materialebestilling.

Der påstås side 7, at den usikkerhed der knytter sig til målingerne er $\pm 3.5\%$

Heraf fremgår at TV-spot var udslagsgivende for p = 34%

1) Beregn et 95% konfidensinterval for p, og kommenter ovennævnte påstand.
2) Hvor mange personer skulle have indsendt spørgeskemaet, hvis påstanden om de 3.5% skulle være korrekt i selv det værst tænkelige tilfælde?

Opgave 6.19
I en analyse af arbejdsgivernes tilfredshed med jobnet, svarede 488 arbejdsgivere på spørgsmålet. Det viste sig, at kun 5% var utilfredse med jobnet.
1) Beregn et 95% konfidensinterval for $p = 0.05$.
2) Giv et skøn over hvor mange arbejdsgivere man skulle have haft svar fra, hvis et 95% konfidensinterval for p skulle have radius 0.01.

Opgave 6.20
I en analyse blev 428 arbejdsgivere spurgt om hvilke jobtyper de annoncerede på jobnet.
Det viste sig, at kun 7% benyttede jobnet til at annoncere efter ledere.
1) Beregn et 95% konfidensinterval for $p = 0.07$
2) Giv et skøn over hvor mange arbejdsgivere man skulle have haft svar fra, hvis et 95% konfidensinterval for p skulle have radius 0.02.

Opgave 6.21
En ny behandling af cancer forventes at give bedre overlevelseschancer end den hidtidige behandling. 120 patienter prøvede den nye behandling, og af disse overlevede 82 i mere end 5 år.
Idet antallet af overlevende patienter antages at være binomialfordelt, skal man
1) Angive et estimat for sandsynligheden p for at overleve i 5 år ved den nye behandling.
2) Angive et 95% konfidensinterval for p.
3) Hvor mange patienter skulle approksimativt lade prøve den nye behandling, hvis radius i 95% konfidensintervallet for p højst skal være 0.05

Opgave 6.22
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 6.23
En fabrikant af lommeregnere er interesseret i at få et skøn over hvor stor en procentdel p af de producerede lommeregnere, der er defekte. En stikprøve på 800 lommeregnere indeholder 10 defekte.
Beregn et 95% konfidensinterval for p.

7. Poissonfordeling

Poissonfordeling benyttes ofte som statistisk model for antallet af "impulser" pr. tidsenhed eller volumenenhed eller længdeenhed o.s.v.
Som eksempler kan nævnes: Antal trafikuheld på en bestemt vejstrækning i løbet af et år, antal revner pr. km kabel, 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.

***SÆTNING* 7.1** *(Poissonfordeling). Lad X angive 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 (μ) med sandsynlighedsfordelingen (tæthedsfunktionen) bestemt ved*

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

Middelværdien for p(μ) er E (X) = μ *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*".*

Et bevis for sætningen føres ikke her

Eksempel 7.1: Poissonfordeling

Der ankommer hver uge (7 dage) i gennemsnit 70 tankskibe med olie til en bestemt havn. Skibene ankommer tilfældigt og uafhængigt af hinanden.
Havnen har kun faciliteter til at modtage højst 15 oliertankskibe om dagen.
a) Hvad er sandsynligheden for at man på en given dag modtager højst 12 tankskibe
b) Hvad er sandsynligheden for, at man på en given dag må afvise olietankere.
Løsning:
Lad *X* betegne antallet af tankskibe der ankommer på en dag. Idet vi med tilnærmelse kan antage, at betingelserne i sætning 7.2 er opfyldt (impuls er her tankskibes ankomst)), er *X* Poissonfordelt *p* (μ). Da det gennemsnitlige antal ankomster pr. dag er $\mu = \frac{70}{7} = 10$ fås:

1) $P(X \le 12) =$ POISSON(12;10;1) = 0,7915
2) $P(X \ge 16) = 1 - P(X \le 15) =$ 1-POISSON(15;10;1) = 0,04874

◆

Konfidensinterval

Som for binomialfordelingen kan man vise, at Poissonfordelingen kan approksimeres med normalfordelingen.
Der gælder:

Konfidensinterval for Poissonfordelt variabel

Lad X værePoissonfordelt p(μ).

Lad der i en stikprøve af størrelsen n være talt m impulser, og lad $\bar{x} = \frac{m}{n}$

Forudsat, at $m \geq 10$ vil $\bar{x}$ være et estimat for μ og et konfidensinterval for μ vil være

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

Eksempel 7.2.Konfidensinterval for Poissonfordeling.

I eksempel 7.1 betragtede vi antallet af tankskibe der anløber en havn.
Ledelsen af havnen føler, at antallet af tankskibe, der anløber havnen er steget, så flere måtte afvises.
Man har derfor planer om at udbygge havnen, men inden da foretager man en optælling af antal skibe der havde anløbet eller ønsket at anløbe havnen i de sidste 30 dage.
Man fandt, at der i alt havde været 360 anløb eller ønsker om anløb på de 30 dage,

1) Angiv på det grundlag et estimat for middelværdien μ af antal anløb pr dag.
2) Angiv et 95% konfidensinterval for μ og angiv på det grundlag, om der var basis for at antage, at antal anløb er steget.

Løsning:

1) På n = 30 dage er der optalt m =360 anløb. Da m>10 kan formel (1) anvendes.
   Vi har $\bar{x} = \frac{360}{30} = \underline{\underline{12}}$
2) Et 95% konfidensinterval for μ er $\bar{x} \pm u_{0.975} \cdot \sqrt{\frac{\bar{x}}{n}} = 12 \pm 1.96 \cdot \sqrt{\frac{12}{30}} = 12 \pm 1.24$.

   [10.76 ; 13.24].
   Da nedre grænse for konfidensintervallet er større end 10, er der basis for at antage, at antallet af tankskibe i middel er steget

Opgaver

Opgave 7.1

Under golfkrigen angreb de allierede flystyrke byen Basra i Irak med i gennemsnit fire angrebsbølger pr. døgn. Angrebsbølgerne blev gennemført uafhængigt af hinanden og på tidspunkter, der ikke fulgte et regelmæssigt skema.

1) Beregn sandsynligheden for, at byen Basra inden for det næste døgn rammes af mindst 5 af de allierede flystyrkers angrebsbølger.
2) Beregn sandsynligheden for, at byen Basra inden for det næste døgn ikke rammes af de allierede flystyrkers angrebsbølger.

Opgave 7.2

En statistik viser, at på en flyvestation har "Brand og Redning" i gennemsnit 3 udrykninger i den første uge af sommerferien. Udrykningerne kommer tilfældigt og uafhængigt af hinanden.

1) Find sandsynligheden for at der højst er 5 udrykninger i den første uge.
2) Statistikken viser også, at "Brand og Redning" i gennemsnit har 5 og 2 udrykninger i 2. og 3. ferieuge.
   Find sandsynligheden for flere end 13 udrykninger i den 3 uger lange sommerferie.

Opgave 7.3

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 pr. 20 m^2 klæde.

Opgave 7.4

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 pr. 10 meter kabel.

Beregn sandsynligheden for, at der

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

Opgave 7.5
Et radioaktivt præparat undergår gennemsnitligt 100 desintegrationer (sønderdelinger) pr. minut. Lad X betegne antal desintegrationer i et sekund (som er lille i forhold til præparatets halveringstid).
Find $P(X \leq 1)$.

Opgave 7.6
Ved en TV-fabrikation optælles som led i en godkendelseskontrol antal loddefejl pr. 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 7.7
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 7.8
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 7.9
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 7.10
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 7.11

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

Idet antallet af isoleringsfejl pr. 50 minutter antages at være Poissonfordelt $p(\mu_1)$, 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 pr. 5 minutter antages at være Poissonfordelt $p(\mu_2)$, skal man

2a) angive et estimat for μ_2.

2b) angive et 95% konfidensinterval for μ_2.

8 EKSPONENTIALFORDELINGEN

I kapitel 7 betragtede man antallet Y af tankskibe, der ankommer til en havn på en dag. Skibene ankommer tilfældigt og uafhængigt af hinanden.

I middel ankom der 10 skibe pr dag, dvs. i middel går der $\lambda = \frac{1}{10}$ dag fra et skib ankommer til det næste skib ankommer.

Vi antog, at Y var Poissonfordelt med middelværdien $\mu = 10$.

Lad X være tidsrummet fra et skib ankommer til det næste ankommer.

Man kan vise (jævnfør sætning 8.1) , at

$$P(X < x) = 1 - e^{-10x} \quad , \qquad P(X > x) = e^{-10x} \quad , \qquad x > 0,$$

Man siger derfor at X er eksponentialfordelt.

Da der i middel går $\lambda = \frac{1}{\mu}$ fra et skib ankommer til det næste ankommer vil man ofte udtrykke eksponentialfordelingen ved sin middelværdi λ , dvs. $P(X < x) = 1 - e^{-\frac{x}{\lambda}}$

Da man sædvanligvis udtrykker en statistisk fordeling ved sin middelværdi er dette sket i følgende sætning hvor vi har ombyttet λ og μ.

Sætning 8.1 Eksponentialfordeling:

Lad Y være en Poissonfordelt stokastisk variabel. Lad det gennemsnitlige antal impulser i en tidsenhed være λ .

I middel går der $\mu = \frac{1}{\lambda}$ tidsenheder mellem 2 impulser.

Lad X = tidsrummet fra en impuls udsendes til den næste udsendes.

X siges at være eksponentialfordelt exp (μ) med parameteren μ, og der gælder

$$P(X < x) = 1 - e^{-\lambda x} = 1 - e^{-\frac{x}{\mu}} \quad , \quad x > 0$$

Middelværdien for exp (μ) er E(X) = μ 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 $P(W = 0) = \frac{(\lambda \cdot x)^0}{0!} \cdot e^{-\mu \cdot x} = e^{-\lambda \cdot x}$, er $P(X > x) = e^{-\lambda x}$.

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

Eksempel 8.1 . Tiden mellem to ankomster.
I kapitel 7 betragtede man antallet Y af tankskibe, der ankommer til en havn på en dag. Skibene ankommer tilfældigt og uafhængigt af hinanden.
I middel ankom der 10 skibe pr dag.
Lad os antage, at en "havnedag" er på 15 timer.
I så fald ankommer der i middel 1.5 skibe pr time.
1) Find sandsynligheden for, at der går mere end 2 timer mellem to på hinanden følgende ankomster.
2) Find sandsynligheden for at der går mellem 1 og 3 timer mellem to på hinanden følgende ankomster

Løsning:
Lad X være tidsrummet fra et skib ankommer til det næste ankommer.
X er eksponentialfordelt med middelværdi $\mu = \frac{1}{1{,}5} = \frac{2}{3}$ time

1) $P(X > 2) = e^{-1.5 \cdot 2} = e^{-3} = \underline{\underline{0.0498}}$

2) $P(1 < X < 3) = P(X < 3) - P(X < 1) = 1 - e^{-1.5 \cdot 3} - \left(1 - e^{-1.5 \cdot 1}\right) = e^{-1.5} - e^{-4.5} = \underline{\underline{0.1929}}$

◆

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:

$$P(X > a) = e^{-\frac{a}{\mu}}, \quad P(X > b) = e^{-\frac{b}{\mu}}, \quad P(X > a+b) = e^{-\frac{a+b}{\mu}} = e^{-\frac{a}{\mu}} e^{-\frac{b}{\mu}}$$

I afsnit 4.4 om betinget sandsynlighed gælder $p(A|B) = \frac{P(A \cap B)}{P(A)}$

Vi har derfor $P(X > a+b | X > a) = \frac{P((X > a+b) \cap (X > a))}{P(X > a)} = \frac{P(X > a+b)}{P(X > a)} = \frac{e^{-\frac{a+b}{\mu}}}{e^{-\frac{b}{\mu}}} = e^{-\frac{b}{\mu}} = P(X > b)$

◆

Eksempel 8.2. 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 μ = 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. $\underline{\underline{30.12\%}}$. ◆

Opgaver

Opgave 8.1

Under golfkrigen angreb de allierede flystyrke byen Basra i Irak med i gennemsnit fire angrebsbølger pr. døgn. Angrebsbølgerne blev gennemført uafhængigt af hinanden og på tidspunkter, der ikke fulgte et regelmæssigt skema.

Lad os antage, at en angrebsbølge er kommet kl 0.00

1) Beregn sandsynligheden for, at den næste angrebsbølge rammer byen Basra inden kl 3.00
2) Find det klokkeslæt t, for hvilken sandsynligheden for, at den næste angrebsbølge kommer mellem kl 0.00 og t er større end 50%.

Opgave 8.2

Ved en øvelse har man etableret et lazaret ved stilling A. Under øvelsen skal lazerettet kun modtage og opbevare "sårede" soldater, der bæres hertil af sanitetstropper på hver sin båre. Øvelsesledelsen har besluttet, at sårede soldater i lazarettet forbliver på båren under resten af øvelsen. I sin planlægning af øvelsen forudser staben, at udkald til bårekrævende sårede soldater vil indtræffe på tilfældige tidspunkter under øvelsen fra kl 8.00 til 12.00 og uafhængigt af hinanden. Den forventede tid mellem to på hinanden følgende bårekrævende udkald antager øvelsesstaben vil være 15 minutter. På lazarettet har man opstillet de 18 bårer, der kan anvendes under øvelsen.

a) Bestem sandsynligheden for, at der højst vil være 20 minutter mellem to bårekrævende udfald.
b) Bestem sandsynligheden for, at der vil være mindst 10 minutter mellem to bårekrævende udfald.
c) Hvad er sandsynligheden for, at der vil være tilstrækkeligt med bårer under øvelsen?
d) Hvad er risikoen for, at lazarettet ikke ser sig i stand til at hente mindst 1 såret soldat som følge af mangel på bårer?
e) Hvor stor er risikoen for, at der efter øvelsens afslutning ligger flere end 5 sårede soldater, som man ikke kunne hente på grund af båremangel.
f) Efter at have vurderet disse tal beslutte øvelsesstaben, at man ikke vil acceptere, at risikoen for, at en såret soldat ikke kan hentes til lazarettet på grund af båremangel, er over 10%. Hvad er det mindste antal bårer, lazarettet efter denne udmelding må prøve at skaffe plads til på lazarettet klar til udkald?

Opgave 8.3

På et betalingsnummer måltes man i tidsrummet fra kl 20 til 22 tiden t (antal minutter) mellem på hinanden følgende telefonopkald. Følgende resultater fandtes:

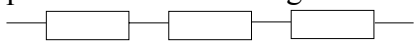
Beliggenhed af t	]0;1]	]1;2]	]2;3]	]3;4]	]4;5]	]5;6]	]6;7]	]7;8]	]8;9]	]9;10]	]10; ∞ [
Antal observationer.	36	21	16	13	7	9	6	1	2	6	0

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 8.4
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 8.5
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 8.6
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?

Tabel 1 Fraktiler u_p i U-fordelingen $n(0,1)$. $P(U \le u_p) = p$.

Bemærk: $u_p = - u_{1-p}$

p	0.0005	0.001	0.005	0.01	0.025	0.05	0.10
u_p	-3.291	-3.090	-2.576	-2.326	-1.960	-1.645	-1.282

p	0.90	0.95	0.975	0.99	0.995	0.999	0.9995
u_p	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Eksempler: $u_{0.975} = 1.960$

Facitliste for udvalgte opgaver

Kapitel 1

1.1 a) kvantitativ, b) kvalitativ c) kvalitativ d)kvalitativ
1.2 -
1.3 -
1.4 -
1.5 (1) - (2) ca 24%
1.6 (1) - (2) ca 0.05

Kapitel 2

2.1 (1) 0.7734 0.0548 0.1718 (2) 0.7480
2.2 (1) 69.15% (2) 10.87% (3) 112.2 (4) 117.3 6.535
2.3 (1) 86.64% (2) 0.008 (3) 0.020
2.4 (1) 5.94% (2) 27.71% (3) [783.51; 816.49]
2.5 1658.76 57.31 3.46%
2.6 103.32 2.6 2.52%
2.7 996.3 46.36 4.65%
2.8 (1) 92.8% (2) 5

Kapitel 3

3.1 (1) 2259.92 35.569 (2) [2237 ; 2283] (3) [2178 ; 2341]
3.2 (1) 74.0362 0.00124 (2) [74.035; 74.037] (3) [74.036 ; 74.037]
3.3 (1) 8.268 0.241 (2) [8.02 ; 8.52] (3) [8.08 ; 8.46]
3.4 [4.23 ; 4.29] 3.5 [0.965 ; 1.111] 3.6 [25.21 ; 60.36]
3.7 [0.00083 ; 0.00231]
3.8 [0.028 ; 0.076]
3.9 [0.0263; 0.1714]

Kapitel 4

4.1 0.1 0.5 0.8 0.2 0.7 1/3
4.2 (1) 0.9134 (2) 0.9678
4.3 (1) 8.75% (2) 38.75% (3) 41.25% (4)11.25%
4.4 (1) 6.4% (2) 78.4% (3) 7.2%
4.5 (1) 27.1% 36.0% 9.756% (2) 53.34% (3) 49.20%
4.6 (1) - (2) 5 3.75 4.082 4.437 (3) 41.67% (4) 12%

Kapitel 5

5.1 (a) 6 (b) 24
5.2 (1) 100 (2) 2400
5.3 3^{40}
5.4 31
5.5 30.24%
5.6 $1.283\ 10^{12}$
5.7 (a) - (b) 736
5.8 60
5.9 90000000

5.10 0.2455
5.11 (A) 0.018% (B) 1.29% (C) 38.24%
5.12 44.57%
5.13 (1) 0.435% (2) 49.57% (3) 41.30%
5.14 (1) 91.67% (2) 25.00% (3) 9.167%
5.15 (1) 17.68% (2) 59.28%

Kapitel 6
6.1 (a) 27.13% (b) 68.46%
6.2 (a) 34.10% (b) 40
6.3 37.11%
6.4 0.0275%
6.5 (a) 9.05% (b) 25
6.6 19.77%
6.7 2.58%
6.8 5.6%
6.9 94.9%
6.10 (1) 2.83% (2) 70.04% (3) 64.97%
6.11 77.86%
6.12 5.83%
6.13 13%
6.14 (1) 7.94% (2) 11.8%
6.15 (1) 60.7% 91.4% 6.9% (2) 4.5%
6.16 40.33%
6.17 (1) 0.108 (2) [0.089 ; 0.127] (3) 21.04
6.18 (1) [0.30; 0.38] (2) ca 784
6.19 (1) [0.03 ; 0.07] (2) ca 1825
6.20 (1) [0.04;0.10] (2) ca 625
6.21 (1) 0.683 (2) [0.600 ; 0.767]
6.22 [0.799 ; 0.847]
6.23 [0.0048 ; 0.0202]

Kapitel 7
7.1 (1) 37.12% (2) 1,83%
7.2 (1) 91.61% (2) 22.8%
7.3 (1) 30.1% (2) 87.9% (3) 4
7.4 (1) 0.292 (2) 0.3482 (3) 0.8965 (4) 7.83
7.5 50.37%
7.6 (1) 15 (2) 81.9%
7.7 75.3%
7.8 6.56%
7.9 76.3%
7.10 44.6%
7.11 (1a) 11 (1b) [4.5 ; 17.5] (2a) 1.1 (2b) [0.45 ; 1.75]

Kapitel 8

8.1 (1) 39.35% (2) 18

8.2 (a) 0.7364 (b) 0.5134 (c) 0.7424 (d) 0.2577 (e) 0.0367 (f) 21

8.3 (1) 2.90 (2) 14.6% (3) 16.98

8.4 (1) 77.88% (2) 10.45% (3) 77.88% (4) 9.48%

8.5 1

8.6 (1) 79.8% (2) 99.33% (3) 22.8 (4) 14

STIKORDSREGISTER