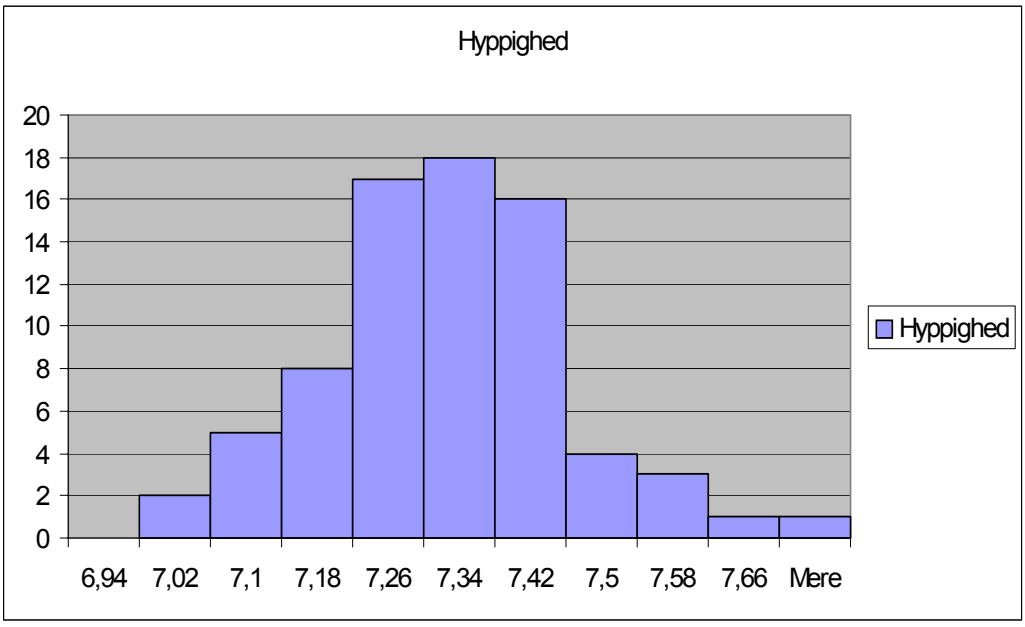
MOGENS ODDERSHEDE LARSEN

ANVENDT STATISTIK

(med anvendelse af Excel)



2. udgave 2008

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.

Da ikke alle de anvendte statistiske funktioner er indbygget fra starten, skal man først vælge et tilføjelsesprogram:
I Excel 2003: Vælg "Funktioner", "Tilføjelsesprogrammer", marker "Problemløser"
I Excel 2007: Vælg "Excel-Office-knappen", "Excel indstillinger (findes forneden)", Tilføjelsesprogrammer","Udfør","marker Problemløser", "Installer".

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

Endvidere forudsættes i enkelte eksempler og opgaver, at data hentes fra min hjemmeside www.larsen-net.dk under "Anvendt statistik".
Endvidere er der til visse eksempler lavet et regneark, hvor mere komplicerede beregninger er foretaget Disse kan også findes på min hjemmeside.

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

marts 2008 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) "inferens" statistik , hvor man ved anvendelse af statistiske metoder søger at slutte (informere) fra en stikprøve til hele proportionen.
   Eksempelvis før et folketingsvalg på basis af en stikprøve på 1000 personer der bliver spurgt om hvem de vil stemme på give en prognose for den forventede mandatfordeling for hele landet (populationen)
   Her vil det være nødvendigt med at kende nogle statistiske metoder til eksempelvis at vide hvor stor en (repræsentativ) stikprøve man skal udtage for at usikkerheden på resultatet er under 5%

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.

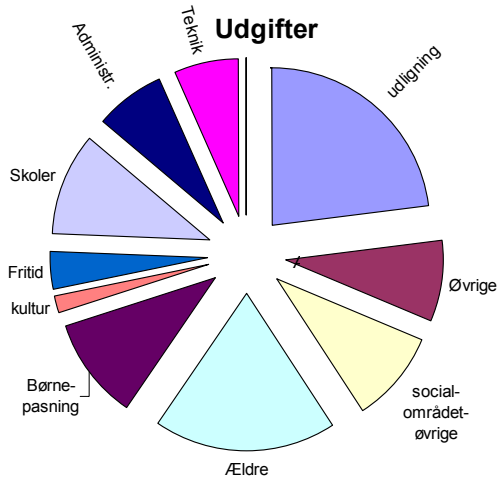
Til illustration af disse data bruges sædvanligvis lagkagediagrammer eller søjlediagrammer

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:

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

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

◆

Eksempel 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

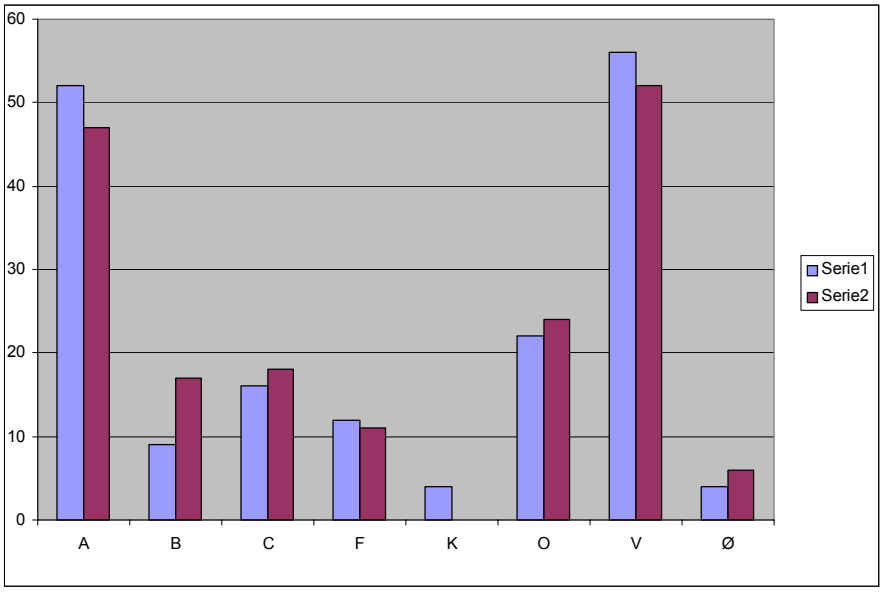
A = Socialdemokraterne, B =Radikale venstre, C = Konservative folkeparti , F =Socialistisk folkeparti, K = Kristendemokraterne, O = Dansk Folkeparti, V = Venstre, Ø = Enhedslisten

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

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

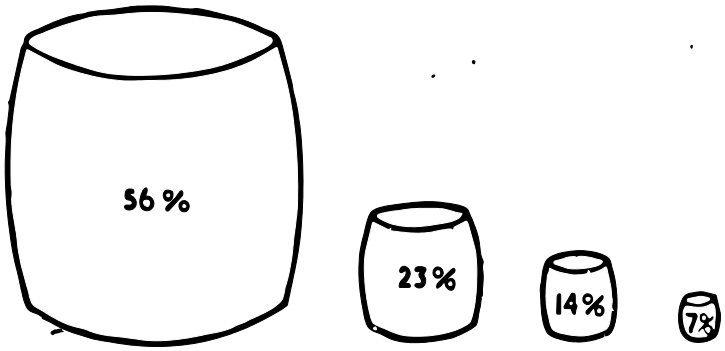
2007: Marker udskriftområde ▶ Vælg på værktøjslinien "Indsæt" ▶ Søjle ▶ Marker ønsket figur



Fordelen ved en grafisk fremstilling er, at de væsentligste egenskaber ved data opnås hurtigt og sikkert. Men netop det, at figurer appellerer umiddelbart til os, gør at vi kan komme til at lægge mere i dem, end det som tallene egentlig kan bære. Eksempelvis viser forsøg, at i lagkagediagrammer, hvor man skal sammenligne vinkler (eller arealer), da vil denne sammenligning afhænge noget af i hvilken retning vinklens ben peger.
Nedenstående eksempel viser hvordan en figur kan være misvisende uden direkte at være forkert.
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 registreres efter det tidspunkt hvor registreringen foregår eller som i eksempel 1.5, hvor det er størrelsen af registrerede 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

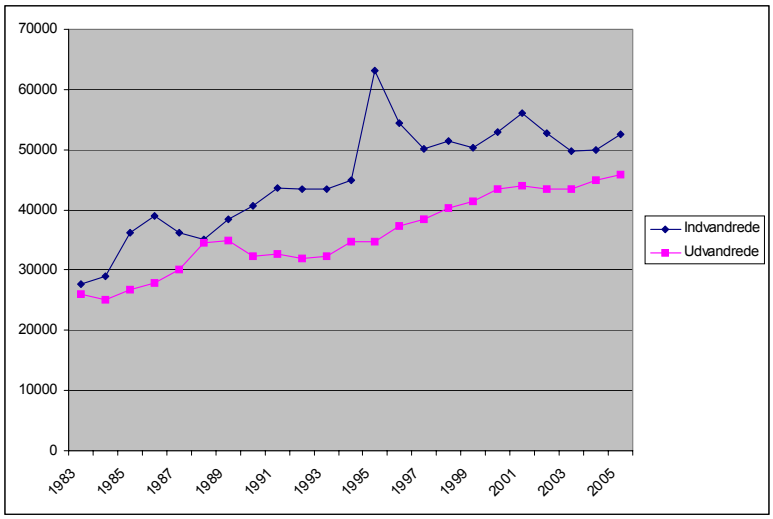
Giv en grafisk beskrivelse af disse data.

Løsning:

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

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

Excel 2007:Marker udskriftsområde ▶ Vælg på værktøjslinien " Indsæt" ▶ Streg ▶ Marker ønsket figur



Eksempel 1.5. Kvantitativ variabel , sideafvigelse ved skydning.
Man har 100 gange målt sideafvigelsen ved skydning med maskingevær.
Resultaterne (som kan findes på adressen www.larsen-net.dk) var følgende:

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

Giv en grafisk beskrivelse af disse data.

Løsning:
I dette tilfælde, hvor vi er interesseret i at få et overblik over tallenes indbyrdes størrelse er det fordelagtigt at tegne et **histogram**.
Et histogram ligner et søjlediagram, men her gælder, at antallet af enheder i hver søjle repræsenteres ved søjlens areal (histo er græsk for areal). Man bør så vidt muligt sørge for at grupperne er lige brede, da antallet af enheder så svarer til højden af søjlen.
Excel 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*
]-24.5 ; -18.7]	//	2
]-18.7 ; -12.9]	/	1
]-12.9 ; - 7.1]	///	3
]-7.1 ; - 1.3]	//////////	11
]-1.3 ; 4.5]	//////////////////	19
]4.5 ; 10.3]	//////////////////////	23
]10.3 ; 15.1]	//////////////////	20
]15.1 ; 21.9]	//////////	12
]21.9 ; 27.7]	///////	7
]27.7 ; 33.5]	//	2

I Excel sker det på følgende måde:

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

2003: Vælg "Funktioner", Dataanalyse, Histogram

2007: Vælg "Data", Dataanalyse, Histogram

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

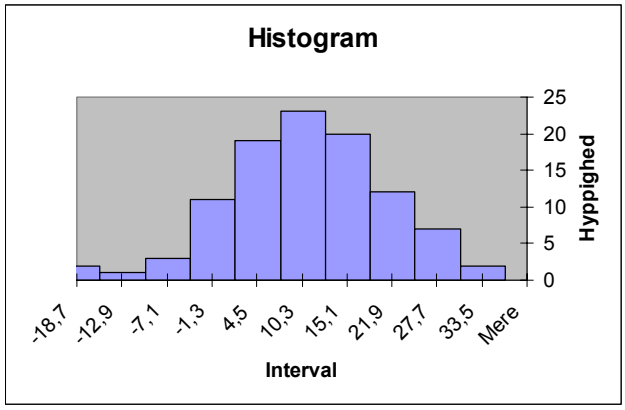
1) Trykkes på OK fås en tabel med hyppigheder, og en figur, hvor intervalgrænserne er fastlagt af Excel.
2) Ønsker man selv at bestemme grænserne, skal man også udfylde intervalområdet. Dette gøres ved at skrive de øvre grænser i en søjle (f.eks. i B1 -18.7, i B2 -12.9 osv.) og så skrive B1:B11 i inputområdet

Nedenstående figurer er blevet gjort lidt "pænere" ved

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

I tilfælde 2 fremkommer følgende

Interval	*Hyppighed*	*Kumulativ %*
-18,7	2	0,00%
-12,9	1	2,00%
-7,1	3	3,00%
-1,3	11	6,00%
4,5	19	17,00%
10,3	23	39,00%
15,1	20	52,00%
21,9	12	79,00%
27,7	7	91,00%
33,5	2	98,00%
Mere	0	100,00%



Det ses, at de fleste målinger ligger fra ca. - 1.3 til ca. 15.1 og så falder hyppigheden nogenlunde symmetrisk til begge sider.

Man regner normalt med, at resultaterne af forsøg, hvor man har foretaget målinger (hvis man lavede nok af dem) har et sådant klokkeformet histogram (beskrives nærmere i kapitel 3)

◆

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.

Vi foretager nu en opsummering(kaldes kumulering), og derefter beregnes ved division med 100 (antal sideafvigelser) tallene i % af det totale antal

Derved fremkommer følgende tabel:

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

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

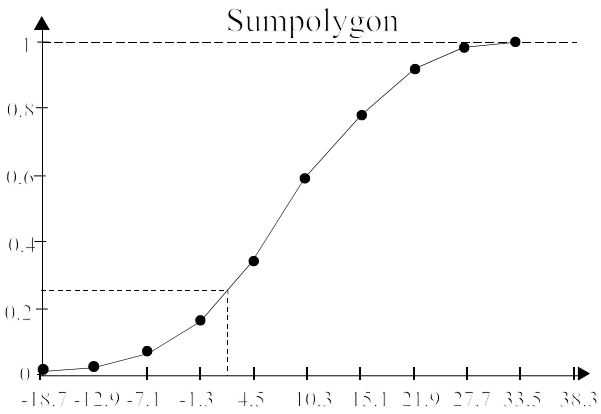
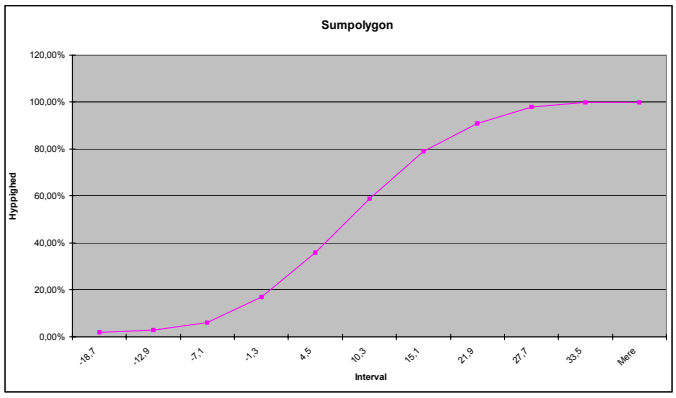


Fig 1.1 *Sumpolygon*

I Excel fås en sumpolygon på følgende måde:
Data indtastes i eksempelvis søjle A1 til A100
2003: Vælg "Funktioner", Dataanalyse, Histogram
2007: Vælg "Data", 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 kumulativ søjle ▶ I værktøjslinien vælges "diagram" ▶ vælg "kurve" osv. ▶ udfør.

Interval	*Hyppighed*	*Kumulativ %*
-18,7	2	2,00%
-12,9	1	3,00%
-7,1	3	6,00%
-1,3	11	17,00%
4,5	19	36,00%
10,3	23	59,00%
15,1	20	79,00%
21,9	12	91,00%
27,7	7	98,00%
33,5	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 populationen består af eller kan inddeles i delmængder (klynger) . Metoden består i, at man ved randomisering vælger et mindre antal klynger, som så totaltælles.

Eksempel: I et vareparti på 2000 emner fordelt på 200 kasser hver med 10 emner ønsker man en vurdering af fejlprocenten.
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*.
2003: Funktioner ▶ Dataanalyse ▶ Beskrivende statistik ▶ udfyld inputområde ▶ Resumestatistik
2007: Data ▶ Dataanalyse ▶ Beskrivende statistik ▶ udfyld inputområde ▶ Resumestatistik
fås følgende udskrift

Kolonne1	
Middelværdi	7,5186
Standardfejl	1,066327
Median	7,26
Tilstand	11,81
Standardafvigelse	10,66327
Stikprøvevarians	113,7053
Kurtosis	0,21644
Skævhed	-0,2137
Område	57,66
Minimum	-24,44
Maksimum	33,22
Sum	751,86
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).

Kaldes observationerne i en stikprøve $x_1, x_2, \ldots, x_n$ er $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$

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 $\bar{x}$ =7.5186.

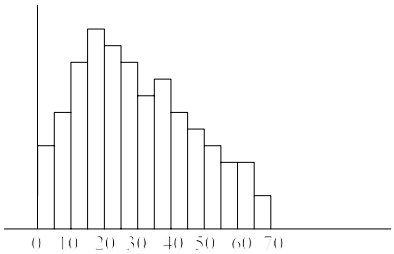
Median: Medianen beregnes på følgende måde:
1) Observationerne ordnes i rækkefølge efter størrelse.
2a) Ved et ulige antal observationer er medianen det midterste tal
2b) Ved et lige antal er medianen gennemsnittet af de to midterste tal.

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

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 7.26, dvs, halvdelen af sideafvigelserne ligger under 7.26.

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

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



Er medianen større end gennemsnittet, 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 kvartil-afstand.** Den viser også, at lønspredningen er væsentlig mindre for række 2 end for række 1 .

◆

I Excel beregnes kvartilerne således:
2003 og 2007: 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 (normalfordelt) 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 $x_1, x_2, \ldots, x_n$

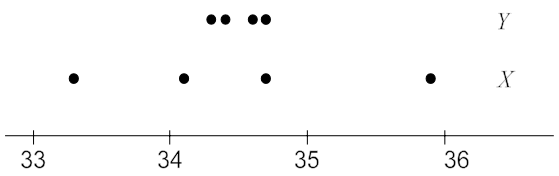
Stikprøvevariansen (eller blot variansen) er s^2 .

Eksempel: Tallene 2,4,5,9 med $\bar{x} = 5$,har variansen $s^2 = \dfrac{(2-5)^2 + (4-5)^2 + (5-5)^2 + (9-5)^2}{4-1} = \dfrac{26}{3} = 8.666$

og spredningen $s = \sqrt{8.667} = 2.94$

Af udskriften i Excel for de 100 værdier fås s =10.66327 og s^2 = 113,7053

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)

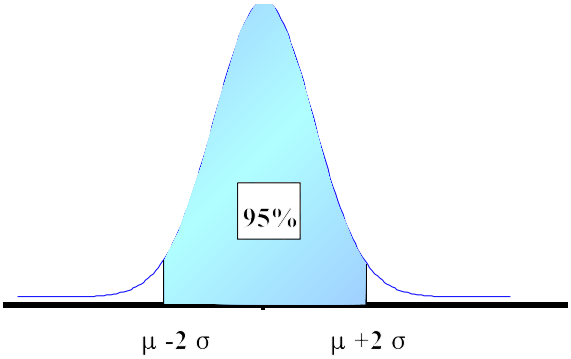


Fig 1.2. *Mellem* $\mu - 2 \cdot \sigma$ *og* $\mu + 2 \cdot \sigma$ *ligger ca.* 95% *af resultaterne.*

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 $\bar{x} = 7.5186$ har en spredning på $s(\bar{x}) = \frac{7.5186}{\sqrt{100}} = 0.7581$

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 0.216 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 - 0.214 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 1.5, 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	$x_i \cdot \frac{n}{100}$
]-24.5 ; -18.7]	- 21.6	2	-0.432
]-18.7 ; -12.9]	- 15.8	1	-0.158
]-12.9 ; -7.1]	- 10.0	3	-0.30
]-7.1 ; - 1.3]	- 4.2	11	-0.462
]-1.3 ; 4.5]	1.6	19	0.304
]4.5 ; 10.3]	7.4	23	1.702
]10.3 ; 15.1]	13.2	20	2.64
]15.1 ; 21.9]	19.0	12	2.28
]21.9 ; 27.7]	24.8	7	1.736
]27.7 ; 33.5]	30.6	2	0.612
]33.5 ; 38.3]	36.4	1	0.364
SUM			8.286

Som det ses er $\bar{x} = 8.286$ tæt ved den "korrekte" værdi 7.518 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 (som kan findes på adressen www.larsen-net.dk) var:

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

1) 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 (som kan findes på adressen www.larsen-net.dk) var (i %):

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

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

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

Opgave 1.7

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

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

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

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..

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 Tætheds- og fordelingsfunktion.

2.1 Indledning

Vi har i kapitel 1 på basis af stikprøver tegnet histogrammer, beregnet gennemsnit og spredning osv. Vi vil i dette afsnit generalisere disse begreber.

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] $\frac{23}{100} = 23\%$

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

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

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

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

Eksempel 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)

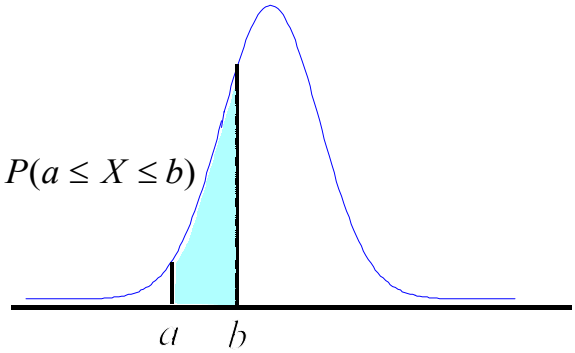


Fig 2.1 *Tæthedsfunktion*

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

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

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

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

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

---

[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]

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

2.3. Fordelingsfunktion

P- fraktil: Lad $f(x)$ være tæthedsfunktionen for en stokastisk variabel X og lad p være et tal mellem 0 og 1.

Ved ***p* - fraktilen** eller $100 \cdot p\%$ fraktilen forstås det tal x_p, for hvilket det gælder, at $P(X \le x_p) = p$ (se figur 2.2)

Medianen m er 50% fraktilen.

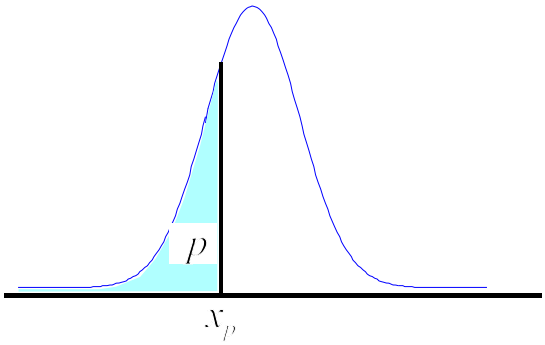


Fig 2.2 *Arealet under kurven til venstre for p - fraktilen x_p er p*

Svarende til at vi tegnede sumkurven i eksempel 1.6 for en stikprøve, kan vi tilsvarende definere en såkaldt **fordelingsfunktionen** $F(x)$ for en stokastisk variabel X.

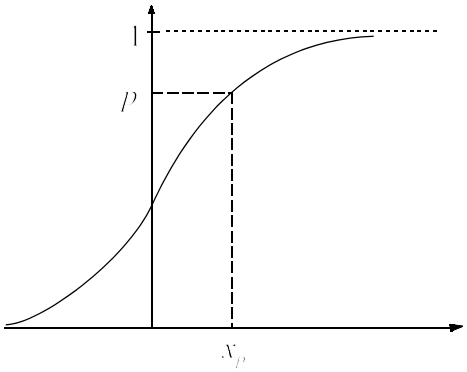


Fig 2.3 *Fordelingsfunktion*

Denne er defineret[3] ved $F(x) = P(X \le x)$

Grafen for $F(x)$ kan ses på figur 2.3

Ved p - fraktilen forstås derfor også det tal x_p for hvilket $F(x_p) = p$

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

2.4. Middelværdi og spredning af sum af stokastiske variable

Statistisk uafhængighed

To hændelser (eller statistiske variable) siges at være uafhængige, hvis sandsynligheden for at den ene hændelse indtræffer ikke afhænger af om den anden indtræffer.

Lad eksempelvis X_1 være resultatet af første kast med en terning og lad X_2 være resultatet ved andet kast med samme terning.
X_1 og X_2 må da anses for at være uafhængige, da resultatet i andet kast ikke er afhængig af hvad resultatet ved første kast.

Middelværdi og spredning af sum af stokastiske variable

Sætning 2.1.

Lad n være et positivt helt tal, lad $X_1, X_2, \ldots, X_n$ være n stokastiske variable, og lad $a_1, a_2, \ldots, a_n$ være n konstanter.

$Y = a_1 \cdot X_1 + a_2 \cdot X_2 + \ldots + a_n X_n$ er da en stokastisk variabel med middelværdien

$E(Y) = a_1 \cdot E(X_1) + a_2 \cdot E(X_2) + \ldots + a_n \cdot E(X_n)$ (sumregel)

Hvis $X_1, X_2, \ldots, X_n$ er statistisk uafhængige, så gælder, at variansen for Y er

$V(Y) = a_1^2 \cdot V(X_1) + a_2^2 \cdot V(X_2) + \ldots + a_n^2 \cdot V(X_n)$ (kvadratregel)

Sætningen bevises ikke her, men illustreres ved følgende eksempel:

Eksempel 2.1 Sum- og kvadratregel

Insektpulver sælges i papkartoner.
Lad X_1 være vægten af insektpulveret og lad X_2 være vægten af papkartonen.
I middel fyldes der 500 gram insektpulver i hver karton. Spredningen på vægten af pulveret er på 5 gram. Kartonerne vejer i middel 10 gram med en spredning på 1.0 gram.

a) Find middelværdien af bruttovægten
b) Find spredningen af bruttovægten.

Løsning.

Bruttovægten er $Y = X_1 + X_2$

a) Ifølge sumreglen gælder da $Y(Y) = E(X_1) + E(X_2) = 500 + 10 = \underline{\underline{510}}$ gram

b) Det synes rimeligt, at vægten af puveret og vægten af papkartonen er uafhængige af hinanden (påfyldningen kan tænkes at ske maskinelt, uden at den på nogen måde er afhængig af hvilken vægt kartonen tilfældigvis har)

Ifølge kvadratreglen gælder da $V(Y) = V(X_1) + V(X_2) = 5^2 + 1^1 = 26$

Spredningen er $\sigma(X) = \sqrt{26} = \underline{\underline{5.1}}$ gram

3 NORMALFORDELINGEN

3.1 Indledning

Ofte vil man finde, at når vi udtager en stikprøve, så vil dens histogram være (næsten) symmetrisk og "klokkeformede". Dette gjaldt eksempelvis for tallene i eksempel 1.5 (sideafvigelse ved skydning). 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.

3.2 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 på figur 3.3 afbildet tæthedsfunktionen for normalfordelingerne $n(0\ ,\ 1)$, $n(4.8\ ,\ 3.2)$, $n(4.8\ ,\ 0.7)$ og $n(10\ ,\ 1)$.

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

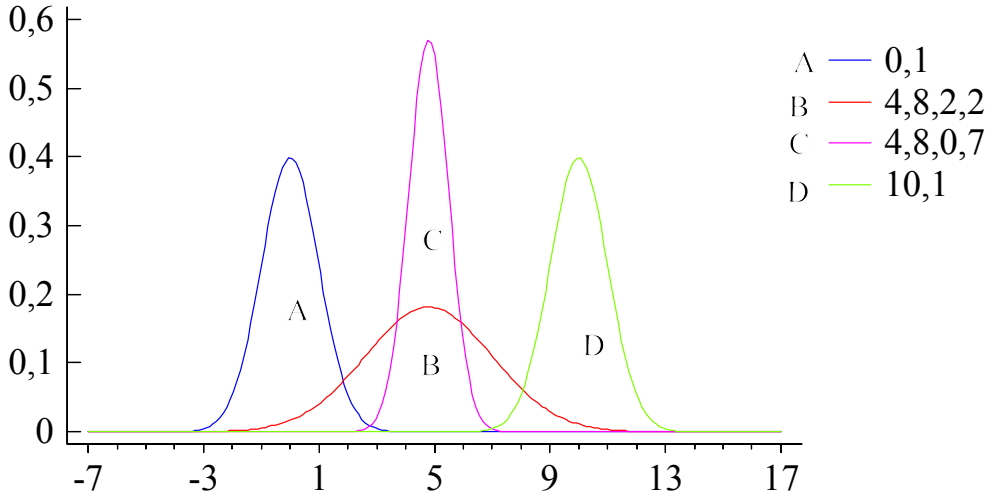


Fig 3.3. *Normalfordelinger med forskellig middelværdi og spredning*

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.

Eksempel 3.2 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.

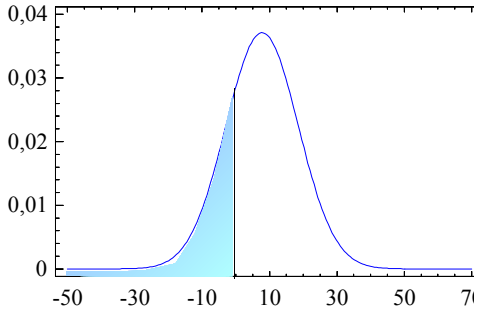
Vi antager nu, at den stokastiske variabel X = sideafvigelserne ved affyring med maskingevær er med tilnærmelse normalfordelt $n(\mu, \sigma)$ med en middelværdi μ = 7.79 og en spredning på σ = 10.75 .

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.

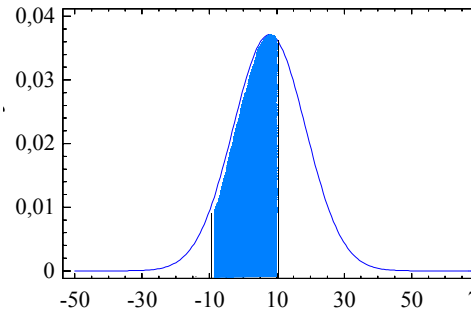
4) Beregn medianen m, dvs. find den sideafvigelse m som halvdelen af skuddene ligger til venstre for. Dette er det samme som at finde m, så $P(X \le m) = 0.50$
5) Beregn 95% fraktilen, dvs. den værdi $x_{0.95}$, som 95% af sideafvigelserne ligger til venstre for. Det er det samme som at sige, at $P(X \le x_{0.95}) = 0.95$

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]

---

[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 $\Phi(u) = P(U \le u) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot \int_{\infty}^{u} e^{-\frac{t^2}{2}} dt$

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

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

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

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

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

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

Løsning:

1) $P(89.2 < X \le 90.8) = P(X \le 90.8) - P(X \le 89.2)$
   =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) ▶
2003: Funktioner ▶ "Målsøgning"
2007: Data ▶ Hvad-hvis analyse ▶ "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(*A*1: *A*12) = 89.97
og s = STDAFV(A1:A2) = 0.435

Vi nævner uden bevis følgende sætning:

Sætning 3.1 Additionssætning.
Lad *n* være et positivt helt tal, lad $X_1, X_2, \ldots, X_n$ være *n* normalfordelte stokastiske variable, og lad $a_1, a_2, \ldots, a_n$ være *n* konstanter.
$Y = a_1 \cdot X_1 + a_2 \cdot X_2 + \ldots + a_n X_n$ er da også en normalfordelt stokastisk variabel

Eksempel 3.4 Additionssætning
En boreproces fremstiller huller med en diameter X_1, der er normalfordelt med en middelværdi μ_1 = 10.00 og en spredning på 0.04. En anden proces fremstiller aksler med en diameter X_2, der er normalfordelt med en middelværdi μ_2 = 9.94 og en spredning på 0.03.
Find sandsynligheden for, at en tilfældig valgt aksel har en mindre diameter end en tilfældig valgt borehul.
Løsning:
$P(X_2 < X_1) = P(X_2 - X_1 < 0)$.
Sættes $Y = X_2 - X_1$ er *Y* normalfordelt.
Ifølge sætning 2.1:sumregel gælder $E(Y) = E(X_2) - E(X_1) = 9.94 - 10.00 = -0.06$.
Da man må antage, at de to processer er uafhængige kan kvadratreglen anvendes
$V(Y) = 1^2 V(X_2) + (-1)^2 V(X_1) = 0.04^2 + 0.03^2 = 0.025$
$\sigma(Y) = \sqrt{0.0025} = 0.05$
$P(X_2 < X_1) = P(Y < 0)$ = NORMFORDELING(0;-0,06;0,05;1) = 0.8849

OPGAVER

Opgave 3.1

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

Opgave 3.2

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 3.3

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

Opgave 3.4

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 3.5

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 3.6
Ved fabrikation af et bestemt mærke opvaskemiddel fyldes vaskepulver i papkartoner. I middel fyldes 4020 g pulver i hver karton, idet der herved er en spredning på 12 g. Pulverfyldningen kan forudsættes ikke at afhænge af kartonernes vægt, der i middel er 250 g med en spredning på 5g.
1) Find spredningen på bruttovægten
2) Beregn sandsynligheden *p* for, at en tilfældig pakke opvaskemiddel har en bruttovægt mellem 4250 g og 4300 g.

Opgave 3.7
Det antages, at den voksne befolkningen i middel vejer 80 kg med en spredning på 10 kg.
1) Find sandsynligheden for at en voksen person vejer over 100 kg.

På en elevator står, at der højst må være 10 personer i elevatoren. Endvidere oplyses, at vægten ikke må overstige 900 kg.
2) Find sandsynligheden for, at den samlede vægt af 10 voksne personer er mere end 900 kg.

Opgave 3.8
I et laboratorium lægges et nyt gulv. Det forudsættes, at vægten Y der hviler på gulvet, er summen af vægten X_1 af maskiner og apparater og vægten X_2 af varer og personale. Da både X_1 og X_2 er sum af mange relativt små vægte, antages det, at de er normalfordelte. Det antages endvidere at X_1 og X_2 er statistisk uafhængige. Erfaringer fra tidligere gør det rimeligt at antage, at der gælder følgende middelværdier og spredninger (målt i tons): $E(X_1) = 6.0,\ \sigma(X_1) = 1.2,\ \mathrm{E}(X_2) = 3.5,\ \sigma(X_2) = 0.4.$
1) Beregn $E(Y)$ og $\sigma(Y)$.
2) Beregn det tal y_0 , som vægten Y med de ovennævnte forudsætninger kun har en sandsynlighed på 1% for at overskride.
3) Beregn sandsynligheden for, at vægten af varer og personale en tilfældig dag, efter at det nye gulv er lagt, er større end vægten af maskiner og apparater. (Vink: se på differensen X_2 - X_1)

4. Konfidensinterval

4.1. 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.

4.2. 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 4.1. 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%◆

4.3. Konfidensinterval for middelværdi

4.3.1. 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 4.1, 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:

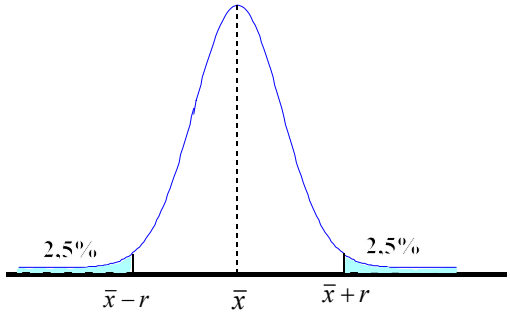


Fig 4.1. *95% konfidensinterval*

Eksempel 4.2 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" $\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{0.5}{\sqrt{12}} = 0.1443$.

Hvis den sande middelværdi μ afviger stærkt fra 90 er det yderst usandsynligt, at vi ville have fået et gennemsnittet på 90.
Eksempelvis, hvis $\mu = 92$ er $P(\bar{X} \le 90) =$ NORMFORDELING(90;92;0,5/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 ;0.1443;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] ◆

[1] I celle A1 skrives en startværdi for μ eksempelvis 90. ▶ I celle B1 skrives =NORMFORDELING(90;A1;0,5/0.1443;1) ▶ Funktioner ▶ "Målsøgning" I "Angiv celle" skrives B1. I "Til Værdi" skrives 0,025. I "Ved ændring af celle" skrives A1. Resultat 90,2841

Som det fremgår af eksempel 4.2 gælder følgende

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

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

Ønskes eksempelvis et 99% eksakt skal $u_{0.975}$ erstattes med $u_{0.995}$ osv.
Disse værdier kan nemmest findes i tabel 1 bagerst i bogen.
Alternativt findes de af Excel: $u_{0.975}$ = NORMINV(0,975;0;1) 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.

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

Eksempel 4.3. 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 99% 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) "Radius" r i et 99% konfidensinterval er $r = u_{0.995} \cdot \frac{\sigma}{\sqrt{n}} = 2.576 \cdot \frac{5.8}{\sqrt{30}} = 3.34$

95% konfidensinterval: $12.65 - 3.34 \le \mu \le 12.65 + 3.34 \Leftrightarrow \underline{\underline{9.31 \le \mu \le 16.00}}$

Vi ved derfor med 95% henholdsvis 99% "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.

4.3.2. 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) (eller formel (2)) blot erstatte σ med s.

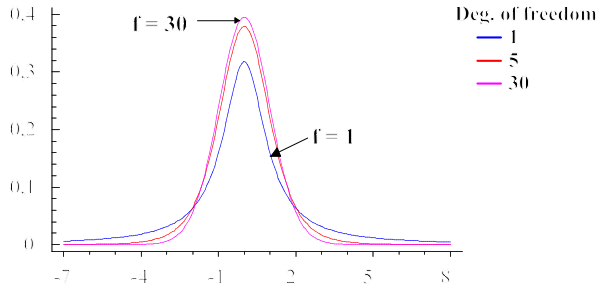
Er stikprøvestørrelsen under 30 bliver denne usikkerhed på s så stor, at man i formel (1) må erstatte U- fraktilen $u_{0.975}$ med en såkaldt t - fraktil $t_{0.975,f}$.

(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 uden praktisk betydning.
Er f lille bliver t - fordelingen så meget bredere end U - fordelingen, at t-fordelingen må anvendes i stedet for U-fordelingen..

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



Ved beregning af konfidensintervaller 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 4.4. 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 $t_{0.975}$, dvs. der beregnes arealet af "øverste hale" hvilket jo også altid er det man har brug for.

$t_{0.975,12}$ = TINV(0.05;12) = 2,178813

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

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

$$\bar{x} - t_{0,975,n-1} \cdot \frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{0.975,n-1} \cdot \frac{s}{\sqrt{n}} \qquad (3)$$

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

Eksempel 4.5. 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) $\bar{x} \pm r = 49.086 \pm r$ hvor r $= t_{1-\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}} = t_{0.975,7-1} \cdot \frac{13.796}{\sqrt{7}}$

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]

Eksempel 4.6 Konfidensinterval, hvis originale data ikke kendt

Find konfidensintervallet for midelværdien μ, idet stikprøven er på 20 tal, som har et gennemsnit på 50 og en spredning på 12.

Løsning:

Nedenstående program findes på adressen www.mogens@larsen-net.dk

	A	B	C	D	E
1	**Eksempel 4.6**		Konfidensradius r =	TINV(B6;B3-1)*B5/KVROD(B3) =	5,616173
2			nedre grænse =	B4-E1	44,38383
3	n =	20	øvre grænse =	B4+E1	55,61617
4	gennemsnit =	50			
5	spredning s =	12			
6	Signifikansniveau α =	0,05			

95% konfidensinterval: [44.38 ; 55.62]

4.3.3. 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 4.7. Dimensionering.
Forstmanden i eksempel 4.4 fandt, at konfidensintervallet der blev beregnet på basis af 7 træer var for bredt.

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

Løsning:

a) På basis af målingerne på de 7 træer sættes $s \approx 14$. Da samtidig $u_{0.975} = 1.96$ fås

$$n = \left(\frac{u_{0.975} \cdot s}{r}\right)^2 = \left(\frac{1.96 \cdot 14}{5}\right)^2 \approx 31$$

Da $n > 30$ er det rimeligt, at benytte en U- fordeling frem for en t-fordeling.
Der skal altså tilfældigt udvælges ca. 32 egetræer.

b) $$n = \left(\frac{u_{0.975} \cdot s}{r}\right)^2 = \left(\frac{1.96 \cdot 14}{6}\right)^2 \approx 21$$

Da $n < 30$ burde man have anvendt en t - fordeling.
Da vi foreløbig regner med 21 træer, erstattes $u_{0.975}$ med $t_{0.975,20} = \text{TINV}(0{,}05;20) = 2.086$.

$$n = \left(\frac{t_{0.975,20} \cdot s}{r}\right)^2 = \left(\frac{2.09 \cdot 14}{6}\right)^2 = 23.8 \approx 24$$

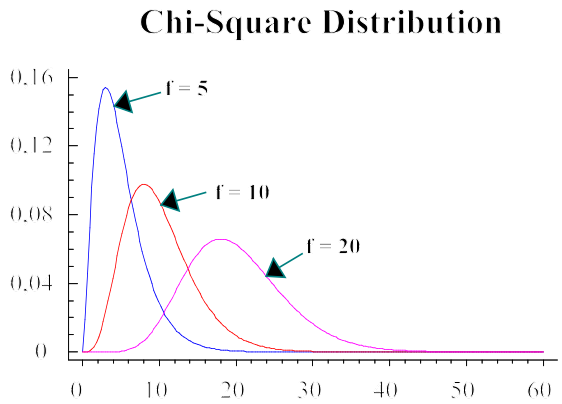
Der skal altså tilfældigt udvælges ca. 24 egetræer.

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

4.4. 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 4.8. 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 4.1

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 4.2

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 4.3

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 4.4

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 4.5

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

108	102	107	98

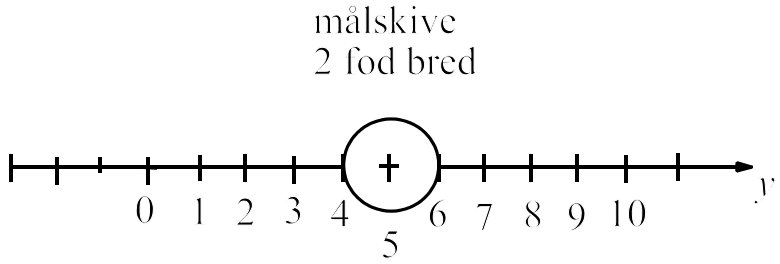
Opstil et 95% konfidensinterval for middelværdien af personens alkoholkoncentration.

Opgave 4.6

Den amerikanske hær har testet et nyudviklet prototype af et gevær, som i et skud kan affyre 1100 "flechettes" (små kugler med små finner i den ene ende til stabilisere dem). I det følgende kort kaldt kugler.

I testen blev prototypen affyret fra 1500 fods afstand mod en målskive, der var 2 fod bred.

Man målte nu hvor hver af de 1100 kugler landede, og afsatte det på en vandret tallinie som vist på figuren. Eksempelvis vil en kugle med en vandret værdi på y = 5.5 ramme skiven, mens en kugle på y = 2.0 ikke rammer skiven.



Ved at ændre indstillingerne på geværet kan man variere spredningen.

PÅ adressen www.larsen-net.dk kan findes en excelfil indeholder resultaterne af 3 skud med spredningen indstillet på ca. 1 fod, 2 fod og 4 fod (benævnt henholdsvis S1, S2 og S4) .

Man er interesseret at finde ud af hvordan en ændring af geværets spredning ændrer antallet af kugler der rammer i en given afstand fra skiven.

Vi betragter i det følgende det skud, hvor man har sat spredningen til 2 fod

1) Find antallet af kugler, der
   a) rammer målskiven, dvs. indenfor intervallet $4 < y < 6$.
   b) rammer indenfor intervallerne $9 < y < 11$ og $-1 < y < 1$.
2) Vurder ved anvendelse af histogrammer og passende karakteristiske tal om fordelingen af skud er approksimativt normalfordelt.
3) Vurder ved beregning af et passende 95% konfidensinterval om påstanden om, at man sigter på skivens midtpunkt på 5 (middelværdien er 5) kan være forkert.

I det følgende antages at fordelingen er normalfordelt med middelværdi 5 og spredning 2.

4a) Find sandsynligheden for, at en kugle rammer målskiven, dvs. indenfor intervallet $4 < y < 6$. Beregn på grundlag heraf, hvor mange kugler, der kan forventes at rammme målskiven.

4b) Find tilsvarende antal kugler, der kan forventes ramme indenfor intervallet $9 < y < 11$.

I det følgende antages, at man kan anvende normalfordelingsapproksimationen for forskellige værdier af spredningen.

5) Beregn for en spredning på 6 det antal kugler, der kan forventes at ramme målskiven og tilsvarende forventes, at ramme indenfor intervallet $9 < y < 11$.

Opgave 4.7.
En nedlagt kemisk fabrik havde muligvis bevirket, at fiskene i en nærliggende flod havde fået et stort indhold af DDT. Specielt var man nervøs for, at sådanne fisk skulle svømme ind i et stort naturreservoir og her forgifte de dyr, som fanger og spiser fisk.
For at undersøge dette nærmere, fangede man fisk på 3 forskellige lokationer, TR, FC, LC og SC hvor TR lå længst væk fra det truede natureservoir.
Man bestemte art (maller, aborrer og karper), vægt (i gram), længde (i centimeter) og DDT niveau (i ppm).
Dataene kan findes på adressen www.larsen-net.dk

1) Beregn hvor mange procent af de fangede fisk der er af hver af de 3 arter, og illustrer tallene ved et passende diagram
2) Beregn tilsvarende hvor mange procent af hver art der blev fanget på hver af de 4 lokaliteter. og begrund, hvorfor man i det følgende koncentrerer sig om at studere mallerne nærmere.
3) Tegn et lagkagediagram over mallernes procentvise fordeling på de 4 lokaliteter.
4) Tegn 3 histogrammer over henholdsvis længde, vægt og DDT-indhold af mallerne, og foretag en vurdering af hvilke af dem der kan tænkes at være normalfordelte.
5) Idet det antages, at vægten for malleren er tilnærmelsesvis normalfordelt, skal man beregne et 95% konfidensinterval for vægten.
6) En af DDt-værdierne for mallerne er en tydelig "outliers", dvs en værdi, som afviger så meget fra de øvrige, at man må antage, at det er en fejlmåling.
   Fjern denne fra talmaterialet, og vurder igen om dette har bevirket at talmaterialet bliver mere normalfordelt.

Opgave 4.8 = opgave 4.1 fortsat.
Find et 95% konfidensinterval for trykstyrkens spredning.

Opgave 4.9 = opgave 4.2 fortsat.
Find ud fra stikprøven et 99% konfidensinterval for diameterens spredning.

Opgave 4.10 = opgave 4.4 fortsat.
Find et 95% konfidensinterval for spredningen af papirets vægt.

Opgave 4.11 = opgave 4.5 fortsat.
Opstil et 95% konfidensinterval for spredningen af personens alkoholkoncentration.

5. Sandsynlighedsregning

5.1 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.

5.2. 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 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 5.1. 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 $P(A) = 0.5$. ◆

5.3 Regneregler for sandsynligheder

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

Eksempel 5.2. Gennemgående eksempel.
To skytter Anders og Brian skyder hver ét skud mod en skydeskive. Sandsynligheden for at Anders rammer skiven er 0.80 mens Brian har en træfsandsynlighed på 0,60.
Et eksperiment består i at de hver skyder et skud.
Lad A være hændelsen at Anders rammer skiven og lad B være sandsynligheden for at Brian rammer skiven.
Vi har derfor, at P(A) = 0.80 og P(B) = 0.60.

◆

Lad os ved at sætte en streg over A forstå "ikke A".
Generelt gælder $\boxed{P(\overline{A}) = 1 - P(A)}$

I eksempel 5.2 er $\overline{A}$ hændelsen at Anders ikke rammer skiven.
Vi har derfor, at P($\overline{A}$) = 1 - P(A) = 1 - 0.8 = 0.20

Fællesmængden til A og B benævnes $A \cap B$ og er mængden af alle udfald i udfaldsrummet U, der tilhører **både A og B** (Den skraverede mængde i figur 5.1).

Eksempelvis er $A \cap B$ i eksempel 5.2 hændelsen, **at både Anders og Brian** rammer skiven

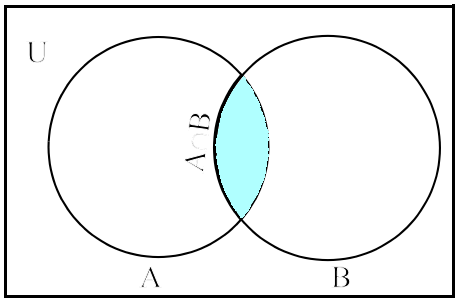


Fig 5.1. *Fællesmængde*

Foreningsmængden af A og B benævnes $A \cup B$ og er mængden af alle udfald i udfaldsrummet U, der **enten tilhører A eller B eventuelt dem begge** (den skraverede mængde på figur 5.2)

Eksempelvis er $A \cup B$ i eksempel 5.2 den hændelse, at enten rammer Anders eller også rammer Brian skiven eventuelt gør de det begge.
Man kunne også udtrykke det ved at mindst en af dem rammer skiven.

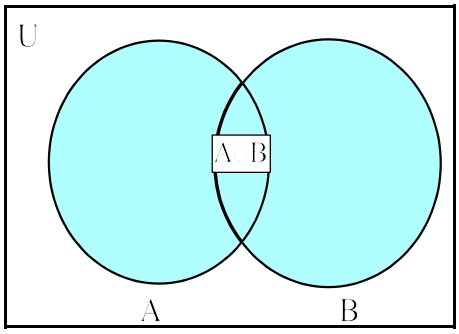


Fig. 5.2 *Foreningsmængde*

Der gælder nu følgende sætninger:

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

Sætningen fremgår umiddelbart ved at betragte arealerne i figur 5.4.

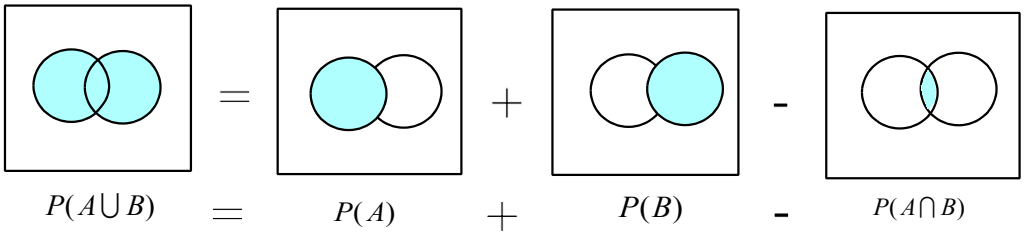


Fig.5.3 *Additionssætning*

Statistisk uafhængighed.
Vi har tidligere nævnt begrebet.
To hændelser A og B siges at være statistisk uafhængige, såfremt sandsynligheden for, at den ene hændelse indtræffer, ikke afhænger af, om den anden hændelse indtræffer.
I eksempel 5.2 må man eksempelvis antage, om Anders rammer skiven har ingen indflydelse på om Brian rammer, så her må man antage A og B er uafhængige.

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

Der gælder følgende sætning:

Produktsætning for uafhængige hændelser:

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

Eksempel 5.3 (eksempel 5.2 fortsat)
Lad A være hændelsen at Anders rammer skiven og lad B være sandsynligheden for at Brian rammer skiven. Det er givet, at P(A) = 0.80 og P(B) = 0.60.
Find sandsynligheden for
a) At både Anders og Brian rammer skiven
b) At enten Anders eller Brian (evt. begge) rammer skiven, dvs mindst en af dem rammer skiven.
c) At hverken Anders elle Brian rammer skiven

Løsning:
a) Da hændelserne antages at være uafhængige gælder ifølge produktsætningen $P(A \cap B) = 0.8 \cdot 0.6 = \underline{\underline{0.48}}$

b) Ifølge additionssætningen gælder $P(A \cup B) = 0.6 + 0.8 - 0.48 = \underline{\underline{0.92}}$

c) $P(\overline{A} \cap \overline{B}) = P(\overline{A}) \cdot P(\overline{B}) = (1-0.8)(1-0.6) = \underline{\underline{0.08}}$ ♦

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

Er hændelserne A og B ikke uafhængige, kan man som beskrevet i afsnit 5.4 udlede en mere generel produktsætning

5.4. Betinget sandsynlighed

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

Eksempel 5.4. Ikke uafhængige hændelser
En fabrik har erfaring for, at den daglige produktion af glasfigurer indeholder 10 % misfarvede, 20% har ridser, og 1 % af produktionen er både ridsede og misfarvede.
Et eksperiment består i tilfældigt at udtage en glasfigur af produktionen. Lad A være hændelsen at få en misfarvet og lad B være hændelsen at få en ridset.
Her er $P(A) \cdot P(B) = 0.1 \cdot 0.2 = 0.02 \neq P(A \cap B) = 0.01$. ♦

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

For at forklare den følgende definition, vil vi simplificere eksempel 5.4, idet vi antager, at den daglige produktion er 100 glasfigurer. I så fald er der 10 misfarvede figurer, 20 ridsede figurer, og 1 figur der er både misfarvet og ridset.
Hvis vi begrænser vort udfaldsrum til A, så er

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

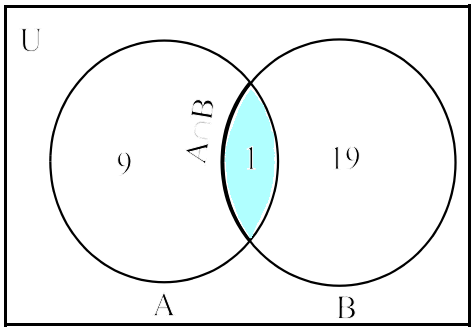


Fig. 5.4 *Taleksempel*

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 $P(B|A) = \dfrac{P(A \cap B)}{P(A)}$.

Ved multiplikation fås

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

Benyttes produktsætningen på eksempel 5.2 fås $P(A \cap B) = P(A) \cdot P(B|A) = 0.1 \cdot 0.1 = 0.01$.

Eksempel 5.4: Betinget sandsynlighed.
En beholder indeholder 3 røde og 3 hvide kugler. Vi udtrækker successivt 2 kugler fra urnen.
Vi betragter følgende 2 hændelser:
A: Den først udtrukne kugle er rød. B: Den anden udtrukne kugle er rød.
Beregn $P(A \cap B)$ hvis
1) kugleudtrækningen foregår, ved at den først udtrukne kugle lægges tilbage før den anden udtrækkes.
2) kugleudtrækningen foregår, ved at den først udtrukne kugle **ikke** lægges tilbage før den anden udtrækkes.
Løsning
1) Her er $P(B|A) = \frac{3}{6}$ og derfor ifølge produktsætningen $P(A \cap B) = P(A) \cdot P(B|A) = \frac{1}{4}$
2) Her er $P(B|A) = \frac{2}{5}$ og derfor $P(A \cap B) = \frac{3}{6} \cdot \frac{2}{5} = \underline{\underline{\frac{1}{5}}}$

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 5.5 (Bayes sætning)
I en officeruddannelse kan man vælge mellem en "teknisk" linie og en "operativ"linie. På en bestemt årgang har 60 % valgt den operative linie og af disse er 20% kvinder. På den tekniske linie er 10% kvinder.
Ved lodtrækning vælges en elev.
a) Find sandsynligheden for, at denne er en kvinde.
Ved ovenstående lodtrækning viste det sig at eleven var en kvinde.
b) Hvad er sandsynligheden for, at hun kommer fra den tekniske linie.
Løsning:
Vi definerer følgende hændelser:
T: Den udtrukne er tekniker
K: Den udtrukne er en kvinde.
a) $P(K) = P(T \cap K) + P(O \cap K) = P(K|T) \cdot P(T) + P(K|O) \cdot P(O) = 0.1 \cdot 0.4 + 0.2 \cdot 0.6 = \underline{\underline{0.16 = 16\%}}$

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 $P(T|K) = \underline{\underline{\dfrac{4}{16} = 25\%}}$

Opgaver

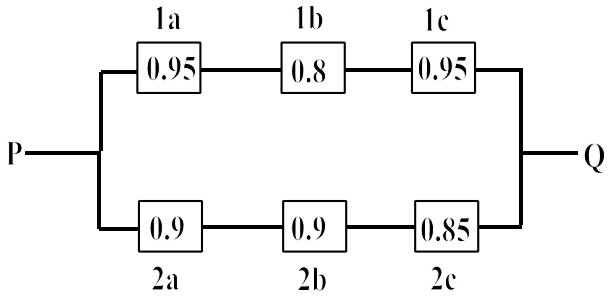
Opgave 5.1

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.

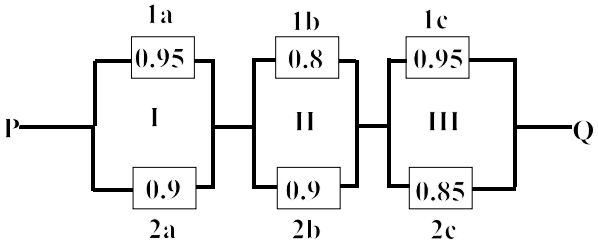
Opgave 5.2

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 5.3
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 5.4
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 5.5
Fire projektgrupper på en virksomhed antages at have sandsynlighederne 0.6, 0.7, 0.8 og 0.9 for at få succes med deres projekt. Grupperne antages at arbejde uafhængigt af hinanden. Find sandsynligheden for, at
a) alle grupper får succes,
b) ingen grupper får succes,
c) mindst 1 gruppe får succes,
d) i alt netop 1 gruppe får succes,
e) i alt netop 3 grupper får succes,
f) i alt netop 2 grupper får succes.

Opgave 5.6
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 5.7

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.

6. Kombinatorik

6.1. 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**.

6.2. 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 6.1. Multiplikationsprincippet

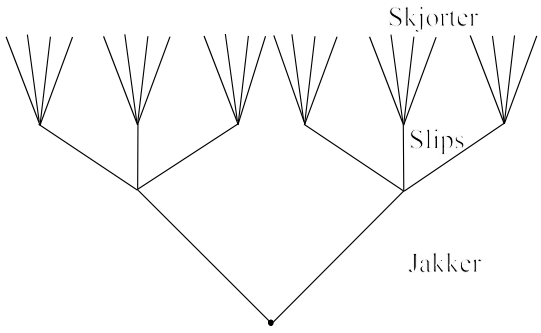
En mand ejer 2 forskellige jakker, 3 slips og 4 forskellige fabrikater skjorter.
På hvor mange forskellige måder kan han sammensætte sin påklædning af jakke, slips og skjorte.

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"



Eksempel 5.2 Fakultet

På hvor mange måder kan 5 personer opstilles i en kø (i rækkefølge)

Løsning:

Pladserne i køen nummereres 1,2,3,4,5.

Plads nr. 1 i køen besættes 5 valgmuligheder
Plads nr. 2 i køen besættes 4 valgmuligheder
Plads nr. 3 i køen besættes 3 valgmuligheder
Plads nr. 4 i køen besættes 2 valgmuligheder
Plads nr. 5 i køen besættes 1 valgmulighed

I alt $5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$ forskellige rækkefølger.

Excel: FAKULTET(5) = 120

Ved ***n* fakultet (*n* udråbstegn) forstås** $n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1$

Endvidere defineres $0! = 1$.

6.3 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

6.3.1 Uden tilbagelægning

Eksempel 6.2. 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 6.2 begrunder følgende definition

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

6.3.2 Med tilbagelægning

Eksempel 6.3. 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 multiplikationsprincippet)	$10 \cdot 10 \cdot 10 \cdot 10 = 10^4$

♦

6.4. Uordnet stikprøveudtagelse

Eksempel 6.4 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 = $\underline{\underline{10}}$

♦

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 6.1 (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 6.5. 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 ♦

6.5 Hypergeometrisk fordeling

Af særlig interesse er den såkaldte "hypergeometriske fordeling", som bl.a. finder anvendelse ved kvalitetskontrol af varepartier (jævnfør eksempel 6.7),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 6.6. 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.

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

Excel: $P(X = 1)$ = KOMBIN(5;1)*KOMBIN(8;3)/KOMBIN(13;4) =$\underline{\underline{0.3916}}$ ♦

Karakteristisk for de såkaldte hypergeometriske fordeling er, at elementerne i udfaldsrummet (kugler i en beholder) kan opdeles i to grupper
En opdeling kunne som i eksempel 5.7 være kvinder og mænd eller som i eksempel 5.8 i defekte elementer og ikke-defekte elementer.
Lad os antage, at vi har en beholder med N kugler, hvoraf de M er røde og resten har en anden farve.
Der udtrækkes en stikprøve på n kugler uden tilbagelægning.
Vi ønsker nu at finde sandsynligheden for at netop x kugler er røde blandt de n udtrukne kugler.
På samme måde som i eksempel 5.7 fås, at denne sandsynlighed bestemmes af formlen

$$P(X=x)=\frac{K(M,x)\cdot K(N-M,n-x)}{K(N,n)}$$

Sætte x = 0, 1, 2, ... finder vi forskellige værdier af $P(X=x)$
Denne sandsynlighedsfordeling kaldes **den hypergeometriske fordeling** med parametrene M,N,N
I Excel behøver man ikke at benytte formlen, da der er indbygget en funktion
HYPGEOFORDELING(x,M,n,N)

Eksempel 6.6 (fortsat)

I eksempel 6.6 var det kvinderne der svarer til antal røde kugler i formen.
Antallet af kandidater var N = 13, hvoraf M = 5 var kvinder. Man udtog en bestyrelse på n = 4.
Man skulle finde sandsynligheden for at netop x = 1 blev udtaget..
Excel:
På værktøjslinien foroven: Tryk på f_x ▶ Vælg kategorien "Statistisk" ▶ Vælg "Hypergeo" ▶ udfylde menuen
. $P(X=1)$ = HYPGEOFORDELING(1;5;4;13) = 0,391608 = 39.16% ◆

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

Eksempel 6.7. 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: $p = P(X=0) + P(X=1)$.
Excel: HYPGEOFORDELING(0;10;25;600)+HYPGEOFORDELING(1;10;25;600)=0,938876
p = 93.89%
Anden beregningsmetode:

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

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

Opgaver

Opgave 6.1.

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

Opgave 6.2.

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 6.3

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 6.4

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 6.5.

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

Opgave 6.6.

I en kasse med 20 elektriske pærer er 5 af pærene sprunget.
8 pærer udtages af kassen uden tilbagelægning. Find sandsynligheden for at højst 1 af disse pærer er sprunget.

Opgave nr. 6.7

Supermarkeder må ikke sælge alkohol til mindreårige. En ekspedient beder blandt unge kunder stikprøvevis halvdelen om ID-identifikation. Blandt 10 unge kunder beder han således de 5 om identifikation.
Hvis der blandt 10 unge kunder er 4 mindreårige, hvad er så sandsynligheden for, at han:
1) Finder alle 4 mindreårige
2) Finder højst 2 mindreårige

Opgave 6.8

En forretning har på lager 100 tyverialarmer, hvoraf de 6 er defekte.
3 tyverialarmer sælges til en kunde. Find sandsynligheden for, at kunden vil få mindst 1 defekt tyverialarm.

Opgave 6.9
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 6.10
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.

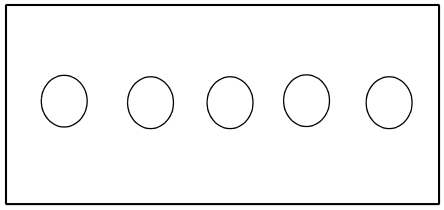
Opgave 6.11
Ved en lodtrækning fordeles 3 gevinster blandt 25 lodsedler. En spiller har købt 5 lodsedler. Beregn sandsynligheden for hver af følgende hændelser:
1) Spilleren vinder alle tre gevinster.
2) Spilleren vinder ingen gevinster.
3) Spilleren vinder netop én gevinst.

Opgave 6.12
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 6.13
En fabrikant fremstiller en bestemt type radiokomponenter. Disse leveres i æsker med 30 komponenter i hver æske. En køber har den aftale med fabrikanten, at hvis en æske indeholder 4 defekte komponenter eller derover, kan køberen returnere æsken, i modsat fald skal den godkendes. Køberen kontrollere hver æske ved en stikprøve, idet han af æsken udtager 10 komponenter tilfældigt. Lad X være antal defekte i stikprøven. Der overvejes nu to planer:
1) Hvis $X = 0$, så godkendes æsken, ellers undersøges æsken nærmere.
2) Hvis $X \leq 1$, så godkendes æsken, ellers undersøges æsken nærmere.
Hvad er sandsynligheden for, at en æske, der indeholder netop 4 defekte komponenter, bliver godkendt af køberen ved metode 1 og ved metode 2.

Opgave 6.14.
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 6.15.
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 6.16
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 6.17
Hvor mange forskellige telefonnumre på 8 cifre kan man danne, når første ciffer ikke må være nul?

Opgave 6.18
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.

7 BINOMIALFORDELING

7.1. Indledning

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

7.2. 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 7.1. En binomialfordelt variabel.
En skytte har 15% sandsynlighed for at ramme målet.
Skytten skyder 6 gange. Hvad er sandsynligheden for at skytten har netop 2 træffere.
Lad X være antallet af træffere blandt de 6 skud
Vi ønsker at finde sandsynligheden for at finde netop 2 træffere blandt disse 6, det vil sige $P(X=2)$.
Løsning:
Lad et eksperiment være at skyde et skud.
Resultatet af eksperimentet har to udfald: træffer , forbier.
Eksperimentet gentages 6 gange uafhængigt af hinanden.
Der er en bestemt sandsynlighed for at få en træffer, nemlig $p = 0.15$.
Lad t være det udfald at få en træffer, og f være det udfald at få en forbier.
Et af de ønskede forløb med 2 træffere vil eksempelvis være *t, f, t, f, f, f, f.*
Dette forløb må have sandsynligheden
$0.15 \cdot (1-0.15) \cdot 0.15 \cdot (1-0.15) \cdot (1-0.15) \cdot (1-0.15) = 0.15^2 \cdot (1-0.15)^4$.
Et andet gunstigt forløb kunne være *f, f, t, f, t, f* med sandsynligheden
$(1-0.15) \cdot (1-0.15) \cdot 0.15 \cdot (1-0.15) \cdot 0.15 \cdot (1-0.15) = 0.15^2 \cdot (1-0.15)^4$
Vi ser, at alle gunstige forløb har samme sandsynlighed.
Antal forløb må være lig antal måder man kan placere to *t*'er på 6 tomme pladser (eller antal måder man kan tage 2 kugler ud af en mængde på 6). Dette ved vi kan gøres på K(6,2) måder.
Vi får følgelig, at $p = K(6,2) \cdot 0.15^2 \cdot (1-0.15)^4 = 0.1762 = \underline{\underline{17.62\%}}$

◆

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

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

Excel: $P(X \le x) = \text{BINOMIALFORDELING}(x, n, p, 1)$
$P(X = x) = \text{BINOMIALFORDELING}(x, n, p, 0)$

Eksempel 7.2 (beregning af binomialfordeling)

I eksempel 7.1 fandt vi, at X var binomialfordelt med n = 6 0g p = 0.15.
$P(X = 2)$ beregnes i Excel på følgende måde:

På værktøjslinien foroven: Tryk på f_x ▶ Vælg kategorien "Statistisk" ▶ Vælg "Binomialfordeling" ▶ udfylde menuen .

Resultat: ==BINOMIALFORDELING(2;6;0,15;0) =0,176177 = 17.6%

◆

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 7.3. Hypergeometrisk fordeling approksimeret med binomialfordeling .

I eksempel 6.7 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 $b\left(25, \frac{10}{600}\right)$.

$P(X \le 1) = \text{BINOMIALFORDELING}(1{,}25{,}1/60{,}1) = \underline{\underline{0.9353}}$

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 $\sigma(p) = \sqrt{\frac{p \cdot (1-p)}{n}}$.

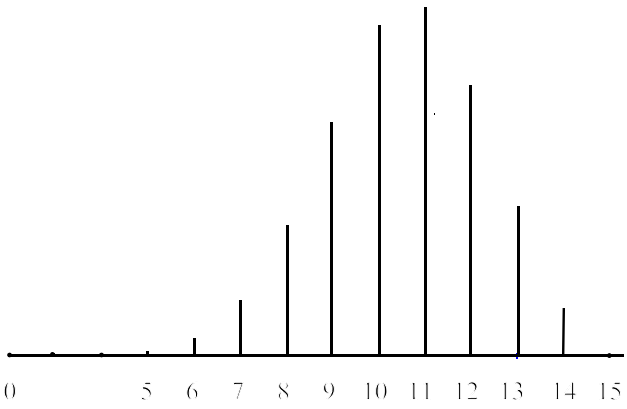
Et bevis vil ikke blive foretaget her.

Eksempel 7.4: 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;0) = 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 $\sigma = \sqrt{n \cdot p \cdot (1-p)} = \sqrt{15 \cdot 0.7 \cdot (1-0.7)} = \underline{\underline{1.77}}$

7.3. 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 7.5. Opinionsundersøgelse.
Ved valget i 2007 stemte 25.5% af vælgerne på socialdemokraterne.
I en opinionsundersøgelse 4 måneder efter valget svarede 1035 vælgere på spørgsmålet om hvilket parti det var mest sandsynligt de ville stemme på hvis der var valg i morgen.
22.7% svarede, at de ville stemme på Socialdemokraterne.
På grundlag heraf blev der konkluderet, at partiet var gået signifikant tilbage siden valget.
Er denne konklusion rimelig?

En metode til at afgøre dette på, er at angive et passende "usikkerhedsinterval" for sandsynligheden p for svarprocenten.

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

Ønsker man at være mere sikker f.eks. 99% sikker, så bliver intervallet naturligvis bredere.
Ofte vil man ved sådanne opinionsundersøgelser vælge 90%.
Da regningerne i princippet er de samme, vil vi i det følgende vælge 95%.

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*

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

I tabel 1 ses, at $u_{0.975} = 1.96$ (eller NormInv(0,975,0,1) = 1,96

Ønskede vi eksempelvis et 90% interval findes tilsvarende, at $u_{0.95} = 1.645$

Eksempel 7.6 Beregning af konfidensinterval

I eksempel 7.5 svarede 1035 vælgere på spørgsmålet om hvilket parti det var mest sandsynligt de ville stemme på hvis der var valg i morgen. 22.7% svarede, at de ville stemme på Socialdemokraterne.

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

Løsning:

Da $\hat{p}$ =0.227 fås $n \cdot \hat{p} \cdot (1-\hat{p}) = 1035 \cdot 0.227 \cdot (1-0.227) \approx 182 \geq 10$

Vi har nu : $0.227 - 1.96 \cdot \sqrt{\frac{0.227 \cdot (1-0.227)}{1035}} \leq p \leq 0.227 - 1.96 \cdot \sqrt{\frac{0.227 \cdot (1-0.227)}{1035}}$

$\Leftrightarrow 0.227 - 0.026 \leq p \leq 0.227 + 0.026 \Leftrightarrow 0.202 \leq p \leq 0.253$

Vi ser altså, at da valgresultatet lå på 25.5%, så må man konkludere, at socialdemokraterne med stor sikkerhed er gået tilbage i forhold til valget.

Excel: (filen kan findes på adressen www.larsen-net.dk)

	A	B	C	D	E	F	
1	p =	0,227			n*p*(1-p)=	181,6125	OK
2	n =	1035		u-værdi =	NORMINV(1-B3/2;0;1)	1,959964	
3	alfa =	0,05		r =	KVROD(B1*(1-B1)/B2)*F2	0,02552	
4				nedre grænse =	b1-f3	0,20148	
5				øvre grænse =	b1+f3	0,25252	

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

Eksempel 7.7. 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 $P(X \leq 10) \approx 0.025$

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 $P(X \geq 10) = 1 - P(X \leq 9) \approx 0.025$

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]

◆

7.4. 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 for et 95% konfidensinterval er $r = u_{0.975} \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}$.

Løses denne ligning med hensyn til n fås

$$n = \left(\frac{u_{0.975}}{r} \right)^2 \hat{p} \cdot (1 - \hat{p})$$

Det grundlæggende problem er her, at man næppe kender $\hat{p}$ eksakt.
Man kender muligvis på basis af tidligere erfaringer størrelsesordenen af $\hat{p}$. Hvis ikke kunne man eventuelt udtage en lille stikprøve, og beregne et $\hat{p}$ på basis heraf.
Endelig er der den mulighed, at sætter $\hat{p} = 0.5$, som er maksimumsværdien af $\hat{p} \cdot (1 - \hat{p})$
Benyttes denne værdi får man den størst mulige værdi af n for en given værdi af r.
Ulempen er, at dette fører til en større stikprøvestørrelse end nødvendigt.
Det følgende eksempel illustrerer fremgangsmåden.

Eksempel 7.8. Dimensionering.
I den i eksempel 7.6 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 7.1
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 7.2
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 7.3
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 7.4
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 7.5
Når SOS har fået anvist erhvervspraktikanter, har det desværre vist sig, at kun 60% af de anviste skoleelever dukker op. Forud for årets praktik har SOS meddelt, at det kun er muligt at gennemføre praktiktjenesten, hvis der dukker mindst 12 praktikanter op.
Skoleelevernes "dukken op" er uafhængig af hinanden.
a) Hvad er sandsynligheden for at SOS kan oprette et hold, hvis praktiktjenesten anviser 15 skoleelever til SOS?
b) Hvor mange elever skal praktiktjenesten anvise, hvis der skal være 95% sandsynlighed for at kurset oprettes.

Opgave 7.6
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 7.7
Antallet af tændstikker i hver af 60 tændstikæsker optaltes.
Resultatet var

43	45	49	47	47	46	46	48	44	48	45	46	42	45	48	47	45	46	50	49
42	44	44	46	44	51	47	46	47	44	41	50	49	44	45	43	50	47	47	45
48	46	48	46	51	48	46	44	49	46	47	49	45	50	46	43	47	43	49	44

Data findes på adressen www.larsen-net.dk .

Lad X være antal tændstikker pr. æske.
1) Tegn et histogram over fordelingen af X.
2) Beregn estimater for middelværdi, spredning og median
3) Vurder ud fra spørgsmål 1 og 2 om X kan antages at være tilnærmelsesvis normalfordelt.
4) Vurder ud fra en hensigtsmæssig konstrueret sumpolygon, hvor stor en procentdel af æskerne, der indeholder mindst 45 tændstikker.
Antag i det følgende at X er normalfordelt, med de i spørgsmål 2 beregnede estimater for middelværdi og spredning.
5) Beregn sandsynligheden for at der i en tilfældig tændstikæske er mindst 45 tændstikker.
Tændstikkerne sælges i pakninger med 10 æsker i hver.
6) Beregn sandsynligheden for, at højst 2 æsker i en tilfældig pakning indeholder mindre end 45 tændstikker.

Opgave 7.8
Blandt familier med 3 børn udvælges 50 familier tilfældigt. Angiv sandsynligheden for, at der i mindst 8 af disse familier udelukkede er børn af samme køn.

Opgave 7.9
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 7.10
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 7.11
Det er oplyst, at der for en given vaccine er 80% sandsynlighed for, at den ved anvendelse har den ønskede virkning.
På et hospital foretoges vaccination af 100 personer med den pågældende vaccine.
Beregn sandsynligheden for, at 15 eller færre af de foretagne vaccinationer er uden virkning.

Opgave 7.12
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 7.13
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 7.14
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 7.15
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 7.16
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 7.17
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 7.18
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 7.19
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.

8. Poisson- og eksponentialfordeling

8.1. Indledning

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

Eksponentialfordelinger benyttes som model for det tidsrum der går fra en impuls udsendes til den næste i ovennævnte Poissonfordeling.

8.2. Poissonfordeling

Lad X angive antallet af "impulser" i et givet tidsrum.
Eksempelvis kunne impulserne være biler der ankommer til en benzinstation pr. time.
Hvis det gennemsnitlige antal impulser i tidsrummet er λ siges X at være Poissonfordelt $p(\lambda)$
Der gælder da, at sandsynligheden for at $X = x$ er bestemt ved

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

Det kan vises, at middelværdien for $p(\lambda)$ er λ og spredningen er $\sqrt{\lambda}$
Et bevis for disse påstande føres ikke her

Eksempel 8.1: 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.

[1] *Ethvert tidspunkt i tidsrummet har samme mulighed for at være impulstidspunkt som ethvert andet tidspunkt. Endvidere skal impulserne indtræffe tilfældigt og uafhængigt af hinanden .*

Løsning:
Lad X betegne antallet af tankskibe der ankommer på en dag. Idet vi med tilnærmelse kan antage, at betingelserne i sætning 8.2 er opfyldt (impuls er her tankskibes ankomst)), er X Poissonfordelt $p(\mu)$. Da det gennemsnitlige antal ankomster pr. dag er $\mu = \frac{70}{7} = 10$ fås:

a) $P(X \le 12) = \text{POISSON}(12;10;1) = \underline{0{,}7915}$

b) $P(X \ge 16) = 1 - P(X \le 15) = 1\text{-POISSON}(15;10;1) = \underline{0{,}04874}$

◆

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

Konfidensinterval for Poissonfordelt variabel
Lad X være Poissonfordelt p(μ).

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

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

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

Eksempel 8.2.Konfidensinterval for Poissonfordeling.
I eksempel 8.1 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

8.3 EKSPONENTIALFORDELINGEN

I afsnit 8.2 betragtede man antallet X 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 $\mu = \frac{1}{10}$ dag fra et skib ankommer til det næste skib ankommer.

Vi antog, at X var Poissonfordelt med middelværdien $\lambda = 10$.

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

Man kan vise (jævnfør sætning 8.1) , at

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

Man siger derfor at T er eksponentialfordelt.

De foregående betragtninger kan udtrykkes ved følgende sætning:

Sætning 8.1 Eksponentialfordeling:

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

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

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

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

Bevis:

I tidsrummet fra t_0 til t_0 + tx er der i gennemsnit $\lambda \cdot t$ impulser.

Lad W være det aktuelle antal impulser i tidsrummet $[t_0 ; t_0 + t\,]$. W er da Poissonfordelt $p(\lambda \cdot t)$.

Idet T er tiden fra én impuls til den næste, er $P(T > t) = P(W = 0)$, da der ingen impulser er i tidsrummet $[t_0 ; t_0 + t\,]$.

Da $P(W = 0) = \frac{(\lambda \cdot t)^0}{0!} \cdot \mathrm{e}^{-\mu \cdot t} = \mathrm{e}^{-\lambda \cdot t}$, er $P(T > t) = \mathrm{e}^{-\lambda t}$.

Vi har derfor $P(T \le t) = 1 - P(T > t) = 1 - \mathrm{e}^{-\lambda \cdot t} = 1 - e^{-\frac{t}{\mu}}$ ◆

Eksempel 8.3. Tiden mellem to ankomster.

I afsnit 8.2 betragtede man antallet X 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 T være tidsrummet fra et skib ankommer til det næste ankommer.

T er eksponentialfordelt med middelværdi $\mu = \frac{1}{1{,}5} = \frac{2}{3}$ time

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

2) $P(1 < T < 3) = P(T < 3) - P(T < 1) = 1 - e^{-1.5 \cdot 3} - \left(1 - e^{-1.5 \cdot 1}\right) = e^{-1.5} - e^{-4.5} = \underline{\underline{0.2120}}$

Excel: 1) $P(T > 2) = 1$ - EKSPFORDELING(2;1,5;1) (bemærk: ikke μ men $\lambda = \frac{1}{\mu}$)

2) $P(1 < T < 3) = P(T < 3) - P(T < 1) =$
EKSPFORDELING(3;1,5;1)-EKSPFORDELING(1;1,5;1) = $\underline{\underline{0{,}212021}}$ ◆

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

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

I afsnit 5.4 om betinget sandsynlighed gælder $p(A|B) = \frac{P(A \cap B)}{P(A)}$

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

Eksempel 8.4. 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.

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 8.2

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 8.3

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 8.4

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

Opgave 8.5

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

Opgave 8.6

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

Opgave 8.7

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

Opgave 8.8

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 8.9

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 8.10

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.

Opgave 8.11

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 8.12

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 8.13

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 8.14

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 8.15

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 8.16

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?

9. Køteori

9.1 Indledning

Oprindelig er køteorien udviklet i forbindelse med dimensionering af telefoncentraler. Senere har teorien fundet anvendelse inden for en lang række forskellige områder som et værktøj der kan anvendes i beslutningsprocesser af økonomisk rækkevidde.
Som eksempler på anvendelser kan nævnes

a) Hvor mange kasser skal der oprettes i et supermarked, så køerne i spidbelastningssituationer ikke bliver for stor.
b) Hvor mange benzinstandere skal en service-station investere i så der ikke er overkapacitet, men heller ikke så få, at mange potentielle kunder opgiver at tanke fordi alle standerne altid er optaget
c) Militær styrker skal i felten råde over et velfungerende kommunikationssystem. Hvordan skal det dimensionere, så eksempelvis i middel højst 10% ikke straks kommer igennem (der er optaget på linien)
d) Hvor mange startbaner skal man bygge i en lufthavn for at man kan betjene trafikken på fuldt betryggende vis.

9.2. En kømodel med M ekspedienter og N pladser i systemet

Vi vil i det følgende antage, at følgende betingelser er opfyldt:

a) Der er $M \geq 1$ ekspedienter i køsystemet
b) Der er i alt N pladser i køsystemet. En kunde, der ankommer til systemet, når der er N kunder i systemet, vil blive afvist.
c) Kundeankomsterne er Poissonfordelte med en gennemsnitlig ankomst på λ kunder pr. tidsenhed.
d) Ekspeditionstiderne er eksponentialfordelt med en gennemsnitlig ekspeditionstid på $\mu = \frac{1}{\alpha}$ tidsenheder pr. kunde. (α er altså det gennemsnitlige antal kunder der ekspedere pr. tidsenhed)
e) En kunde kan frit vælge en ledig ekspedient.

Situationen er anskueliggjort i figur 9.1.

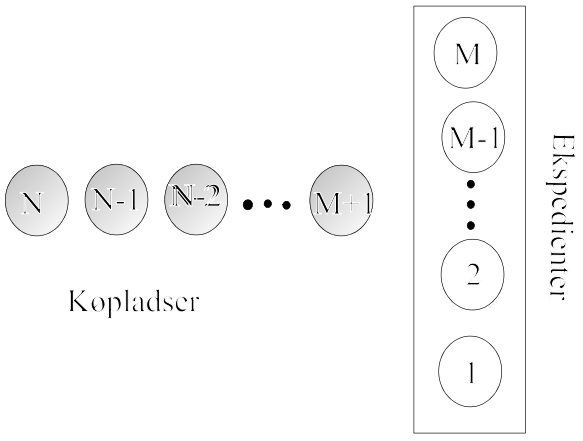


Fig 9.1. *Kømodel*

Et køsystem der opfylder disse betingelser kaldes en "klassisk kømodel".

Eksempel 9.1. Kø ved benzinstation

En benzinstation har 4 benzinstandere , og der er på pladsen plads til 4 biler, der venter på at blive ekspederet. På stationen har man fået gennemført en analyse, der viser, at der i den travleste tid i gennemsnit i et tidsrum på 5 minutter ankommer 2 kunder , og det i gennemsnit tager 5 minutter for en kunde at blive ekspederet (fylde benzin på bilen og betale).
Vi antager at betingelserne for en klassisk kømodel er opfyldt.
Med de ovenfor angivne betegnelser er M = 4 og N = 8, og regnes med en tidsenhed på 1 minut, så er $\lambda = \frac{2}{5}$ og $\mu = 5$. ◆

Et sådant system kan karakteriseres ved en række begreber der defineres i det følgende:

Tilstand E_n :

Man siger, at systemet er i tilstanden E_n hvis der er *n* kunder i systemet.

Tilstandssandsynlighed P_n :

Ved **tilstandssandsynligheden** P_n forstås sandsynligheden for at være i tilstand E_n .

Statistisk ligevægt:

Man siger, at systemet er i **statistisk ligevægt**, hvis tilstandssandsynlighederne er uafhængige af begyndelsestilstanden og af tiden t. Det betyder ikke, at systemet er i samme tilstand (den kan selvfølgelig skifte mellem en tilstand E_n og E_{n+1} eller E_{n-1}, men det skal forstås sådan, at sandsynligheden for at finde systemet i en given tilstand er uafhængig af tiden og af begyndelsestilstanden.
Systemerne vil i praksis altid meget hurtigt komme i statistisk ligevægt, så det er en forudsætning for de efterfølgende formler.

Trafiktilbud: ρ : $\rho = \lambda \cdot \mu = \frac{\lambda}{\alpha}$ (Enheden kaldes Erlang efter dansk matematiker)

Det gennemsnitlige antal kundeankomster pr. middelekspeditionsperiode

Overslagsbetragtninger:

Uden beregninger kan man umiddelbart se, at hvis antallet af kunder i gennemsnit er større end det antal som de M ekspedienter kan ekspedere, dvs. hvis $\frac{\lambda}{\alpha \cdot M} > 1$ så må der være en stor sandsynlighed for at alle ekspeditionssteder er optaget, og at der er mange kunder der må vente og der må ske en del afvisninger.

Da $\frac{\lambda}{\alpha \cdot M} > 1 \Leftrightarrow \frac{\lambda}{\alpha} > M \Leftrightarrow \rho > M$ ses, at hvis trafiktilbuddet $\rho > M$ har vi ovennævnte forhold, mens hvis $\rho < M$ så har vi omvendt, at der i gennemsnit er ledige ekspedienter, ingen kø osv.

I eksempel 9.1 er $\rho = \frac{2}{5} \cdot 5 = 2 < M = 4$ så vi må forvente at der stort set ikke er kø osv.

Man kan nu udlede en række formler, der kan benyttes til at vurdere om systemet har de optimale dimensioner.

Disse udledninger er ret omfattende og vil ikke blive gennemgået her, men resultaterne er samlet i følgende skemaer.

Tabel 9.1a

nr	Tilstand	Sandsynlighed	
1	E_0 : 0 kunder	P_0 : 0 kunder	$\left(\frac{\rho^{M+1}}{M!} \cdot \frac{1-\left(\frac{\rho}{M}\right)^{N-M}}{M-\rho} + \sum_{i=0}^{M} \frac{\rho^i}{i!} \right)^{-1}$ for $\rho \neq M$ $\left((N-M) \cdot \frac{\rho^M}{M!} + \sum_{i=0}^{M} \frac{\rho^i}{i!} \right)^{-1}$ for $\rho = M$
2	E_n : n kunder	P_n : n kunder	$\frac{\rho^n}{n!} \cdot P_0$ for $n \leq M$ $\frac{\rho^M}{M!} \cdot \left(\frac{\rho}{M}\right)^{n-M} \cdot P_0 = \left(\frac{\rho}{M}\right)^{n-M} \cdot p_M$ for $M < n \leq N$ 0 for n > N
3	Systemet er blokeret	P_N	$\frac{\rho^M}{M!} \cdot \left(\frac{\rho}{M}\right)^{N-M} \cdot P_0 = \left(\frac{\rho}{M}\right)^{N-M} \cdot p_M$
4	Alle ekspeditionssteder er optaget, dvs. $E_M, E_{M+1}, \ldots, E_{N-1}$	At blive forsinket F = $\sum_{i=M}^{N-1} P_i$,	$\frac{M}{M-\rho} \cdot (P_M - P_N)$ for $\rho \neq M$ $(N-M) \cdot P_N$ for $\rho = M$
5	Mindst et ekspeditionssted er ledigt	Straks at blive ekspederet S = $\sum_{i=0}^{M-1} P_i$	$1 - F - P_N$

Tabel 9.1b

6	Gennemsnitlig kølængde M < N	G = $\sum_{i=1}^{N-M} i \cdot P_{M+i}$	$\frac{\rho}{M-\rho}(F-(N-M)\cdot P_N)$ for $\rho \neq M$ $\frac{(N-M)\cdot(N-M+1)\cdot P_N}{2}$ for $\rho = M$
7	Gennemsnitlig antal optagne ekspedienter= det gennemsnitlige antal kunder under ekspedition	E= $\sum_{i=0}^{M-1} i \cdot P_M + M \cdot \sum_{i=M}^{N} P_i$	$\rho \cdot (1 - P_N)$
8	Gennemsnitligt antal kunder i systemet	E+G	
9	Middelventetid i køen for alle kunder	$V_{KØ}$	$\frac{G}{\lambda}$
10	Middelventetid i køen for forsinkede kunder	$W_{KØ}$	$\frac{G}{\lambda \cdot F}$
11	Middelventetid i køsystemet (i kø + ekspedition) for forsinkede kunder	W_{SY}	$\frac{G}{\lambda \cdot F} + \mu = W_{KØ} + \mu$

Eksempel 9.2. Fortsættelse af eksempel 9.1

En benzinstation har 4 benzinstandere , og der er på pladsen plads til 4 biler, der venter på at blive ekspederet. På stationen har man fået gennemført en analyse, der viser, at der i den travleste tid i gennemsnit i et tidsrum på 5 minutter ankommer 2 kunder , og det i gennemsnit tager 5 minutter for en kunde at blive ekspederet.

a) Find sandsynligheden for, at der ingen kunder er
b) Find sandsynligheden for at alle 8 pladser er optaget (at en kunde bliver afvist)
c) Find sandsynligheden for at en kunde vil blive forsinket (skulle vente i køen).
c) beregn hvor mange benzinstandere der i gennemsnit er i brug

Løsning:

Med de i tabel 9.1a angivne betegnelser er M = 4 og N = 8.

Regnes med en tidsenhed på 1 minut, så er $\lambda = \frac{2}{5}$ og $\mu = 5$.

Heraf følger, at trafiktilbuddet er $\rho = \lambda \cdot \mu = \frac{2}{5} \cdot 5 = 2$ Erlang

a) Ifølge tabel 9.1a nr. 1 er $P_0 = \left(\frac{\rho^{M+1}}{M!} \cdot \frac{1 - \left(\frac{\rho}{M}\right)^{N-M}}{M - \rho} + \sum_{i=0}^{M} \frac{\rho^i}{i!} \right)^{-1}$

Ved indsættelse fås (lommeregner benyttet)

$$P_0 = \left(\frac{2^{4+1}}{4!} \cdot \frac{1 - \left(\frac{2}{4}\right)^{8-4}}{4-2} + \sum_{i=0}^{4} \frac{2^i}{i!} \right)^{-1} = \left(\frac{2^5}{4!} \cdot \frac{1 - \left(\frac{1}{2}\right)^4}{2} + \frac{2^0}{0!} + \frac{2^1}{1!} + \frac{2^2}{2!} + \frac{2^3}{3!} + \frac{2^4}{4!} \right)^{-1}$$

$$\left(\frac{32}{24} \cdot \frac{1 - \frac{1}{16}}{2} + 1 + 2 + 2 + \frac{8}{6} + \frac{16}{24} \right)^{-1} = \left(\frac{5}{8} + 7 \right)^{-1} = \left(\frac{61}{8} \right)^{-1} = \frac{8}{61} = 0.1311$$

Konklusion: Der er ingen biler på tankområdet i 13.1% af tiden

Da P_0 giver anledning til ret store regninger, og indgår i mange af formlerne er der nedenfor angivet et Excel regneark med beregninger anført (findes som excelfil i www.larsen-net.dk). Anvendes programmet til en anden opgave skal man også huske i celle e2 at ændre B12 til B(8+M)

	A	B	C	D	E
	Eksempel 9.2		led 1	HVIS(B6<>B3;B6^(B3+1)/FAKULTET(B3)*(1-(B6/B3)^(B2-B3))/(B3-B6);(B2-B3)*B6^B3/FAKULTET(B3))	0,625
2	N =	8	sum	SERIESUM(B6;0;1;B8:B12)	7
3	M =	4	p0 =	(E1+E2)^(-1)	0,131148
4	λ =	0,4	pM =	B6^B37/FAKULTET(B3)*E3	0,087432
5	μ =	5	pN =	B6 ^B3/FAKULTET(B3)*(B6/B3)^(B2-B3)*E3	0,005464
6	ρ =	2			
7	Koefficienter				
8		1			
9	1/FAKULTET(1)	1			
10	1/FAKULTET(2)	0,5			
11	1/FAKULTET(3)	0,166667			
12	1/FAKULTET(4)	0,041667			
13	1/FAKULTET(5)	0,008333			

b) Ifølge tabel 9.1a nr. 3 er $P_N = \frac{\rho^M}{M!} \cdot \left(\frac{\rho}{M}\right)^{N-M} \cdot P_0$

Ved indsættelse fås (lommeregner benyttet)

$$P_8 = \frac{2^4}{4!} \cdot \left(\frac{2}{4}\right)^{8-4} \cdot 0.1311 = \frac{16}{24}\left(\frac{1}{2}\right)^4 0.1311 = 0.00546$$

Konklusion: Der er kun fuldt optaget på tankstationen i 0.05% af tiden.
Der er altså megen ringe sandsynlighed for at en kunde afvises.

c) Ifølge tabel 9.1a nr. 4 er F = $\frac{M}{M-\rho} \cdot (P_M - P_N)$

Ved indsættelse fås (lommeregner benyttet)

$$F = \frac{4}{4-2} \cdot (P_4 - P_8) = 2\left(\frac{2^4}{4!} p_o - 0.00546)\right) = 2\left(\frac{16}{24} 0.1311 - 0.00546\right) = 0.164$$

Konklusion: I middel vil 16.3% af kunderne opleve, at de skal vente i køen (blive forsinket)

d) Ifølge tabel 9.1b nr. 7 er E = $\rho(1 - P_N)$
Ved indsættelse fås (lommeregner benyttet): E = 2 (1 - 0.00546) = 1.99
Konklusion: I gennemsnit er ca. 2 benzinstandere i brug

Ud fra disse tal kunne ejeren af tankstationen nu vurdere, om anlægget er fornuftigt dimensioneret.

◆

9.3. Køsystemer med plads til et ubegrænset antal ventende kunder.

I en del tilfælde, er der plads til så mange ventende kunder, at der i praksis er plads til et ubegrænset antal.
Eksempelvis på et stort posthus, hvor der er 4 ekspedienter og man trækker et nummer. Da man jo kan stå meget tæt (eventuelt helt ud på gaden) er køen i praksis ubegrænset.

Idet vi stadig antager at køsystemet opfylder kravene til en klassisk kømodel, så giver tabel 9.2 en oversigt over formler der kan anvendes i dette tilfælde.
Formlerne gælder kun for $\rho < M$
Dette skyldes, at hvis $\rho > M$ vil køen gradvist vokse, og man når derfor aldrig til statistisk stabilitet.(matematisk skyldes det , at man ved udledningen har benyttet nogle regler for uendelige rækker, der så ikke har nogen endelig sum).

Strengt taget er det i nedenstående tabel unødvendigt at se specielt på tilfældet M = 1, da den generelle formel gælder for M ≥ 1 dvs. også for M = 1. men da formlen i dette tilfælde, er blevet meget simpel, så anføres denne særskilt.

Tabel 9.2a

nr	Tilstand	Sandsynlighed	Generel forudsætning $\rho < M$
1	E_0 : 0 kunder	P_0 : 0 kunder	$\left(\frac{\rho^M}{M!}\cdot\frac{\rho}{M-\rho}+\sum_{i=0}^{M}\frac{\rho^i}{i!}\right)^{-1}$ $1-\rho$ for M = 1
2	E_n : n kunder	P_n : n kunder	$\frac{\rho^n}{n!}\cdot P_0$ $n \le M$ $\frac{\rho^M}{M!}\cdot\left(\frac{\rho}{M}\right)^{n-M}\cdot P_0$ $n > M$ $\rho^n \cdot p_0$ for M = 1
4	Alle ekspeditionssteder er optaget, dvs. $E_M, E_{M+1}, \ldots, E_{N-1}$	At blive forsinket F	$\frac{M}{M-\rho}\cdot P_M$ ρ for M = 1
5	Mindst et ekspeditionssted er ledigt	Straks at blive ekspederet S	$1-F$ $1-\rho$ for M = 1

Tabel 9.2b

6	Gennemsnitlig kølængde	G	$\frac{\rho}{M-\rho}\cdot F$ $\frac{\rho^2}{1-\rho}$ for M = 1
7	Gennemsnitlig antal optagne ekspedienter= det gennemsnitlige antal kunder under ekspedition	E	ρ
8	Gennemsnitligt antal kunder i systemet	E+G	
9	Middelventetid i køen for alle kunder (i kø, afviste og straksekspederede)	$V_{KØ}$	$\frac{G}{\lambda}$ $\frac{\rho^2}{\lambda\cdot(1-\rho)}$ for M = 1
10	Middelventetid i køen for forsinkede kunder	$W_{KØ}$	$\frac{G}{\lambda\cdot F}$
11	Middelventetid i køsystemet (i kø + ekspedition) for forsinkede kunder	W_{SY}	$\frac{G}{\lambda\cdot F}+\mu = W_{KØ}+\mu$

Eksempel 9.3 Ubegrænset plads

På et posthus er der 4 ekspedienter, og i praksis ubegrænset plads til ventende kunder.
Kunderne trækker et nummer, og står derfor i en fælles kø til de 4 ekspedienter.
En undersøgelse viser, at hver ekspedient i gennemsnit kan ekspedere en kunde på 2 minutter.
Kunderne ankommer med en gennemsnitlig intensitet på 90 kunder pr. time.
Det antages, at den klassiske kømodel kan anvendes.

a) Find det gennemsnitlig antal optagne ekspedienter
b) Find den gennemsnitlige kølængde
c) Middelventetiden i køsystemet for alle kunder

Løsning:

Man har, at M = 4, $\lambda = \frac{90}{60} = 1.5$, $\mu = 2$ og $\rho = 1.5 \cdot 2 = 3$

Da $\rho < M$ kan de i tabellen nævnte formler benyttes:

a) Det gennemsnitlig antal optagne ekspedienter E = 3

b) Af tabel 9.2 nr. 6 fås G = $\frac{\rho}{M-\rho} F = \frac{3}{4-3} F = 3F = 3\frac{4}{4-3} p_M = 12 \cdot P_M = 12 \cdot \frac{3^4}{4!} P_0$

$$= \frac{81}{2} P_0 = \left(\frac{81}{2} \left(\frac{\rho^M}{M!} \cdot \frac{\rho}{M-\rho} + \frac{\rho^0}{0!} + \frac{\rho^1}{1!} + \frac{\rho^2}{2!} + \frac{\rho^3}{3!} \right) \right)^{-1} = \left(\frac{81}{2} \left(\frac{3^4}{4!} \cdot \frac{3}{4-3} + 1 + 3 + \frac{3^2}{2} + \frac{3^3}{3!} + \frac{4^4}{4!} \right) \right)^{-1}$$

= 1.53

Den gennemsnitlige kølængde er på 1.53 kunder.

c) Middelventetiden i køsystemet for alle kunder

$W_{SY} = \frac{G}{\lambda \cdot F} + \mu$

Af spørgsmål b) ses, at G = 3 F, dvs. $W_{SY} = \frac{3}{\lambda} + \mu = 2 + 2 = 4$

Excel:(findes som excelfil i www.larsen-net.dk)

	A	B	C	D	E
1	**Eksempel 9.3**		led 1	HVIS(B6<B3;B6^B3)/FAKULTET(B3)*B6/(B3-B6);"ro>M")	10,125
2	n =	1	sum	SERIESUM(B6;0;1;B8:B12) Bemærk: B8 til B(8+M)	16,375
3	M =	4	p0 =	(E1+E2)^(-1)	0,037736
4	λ =	1,5	pM =	B6^B3/FAKULTET(B3)*E3	0,127358
5	μ =	2	pn =	HVIS(B2<=B3;B6^B2/FAKULTET(B2)*E3;B6^B2/(B3*FAKULTET(B3))*E3)	0,113208
6	ρ =	3			
7	Koefficienter				
8		1			
9	1/FAKULTET(1)	1			
10	1/FAKULTET(2)	0,5			
11	1/FAKULTET(3)	0,166667			
12	1/FAKULTET(4)	0,041667			
13	1/FAKULTET(5)	0,008333			

Opgaver

Opgave 9.1
I forbindelse med en international militær operation skal en signalofficer vurdere kapaciteten af en radioforbindelse til en underlagt enhed.
Radiosystemet tillader, at 2 samtaler kan stilles i kø.
Signalofficeren forventer et behov for 15 samtaleopkald pr. time og hvert opkald varer i middel 2 minutter.
Anvend den klassiske kømodel til at vurdere om sandsynligheden for at vurdere om sandsynligheden for at samtaleopkald kan afvises kan holdes under 0.1.

Opgave 9.2
En militærlæge har plads til 10 patienter i sit venteværelse. Det antages, at en patent, der finder, at venteværelset er fuldt, går igen (afvises fra systemet).
Patienterne ankommer til lægen med en gennemsnitsværdi på 1.2 patienter pr. 20 minutter.
Lægens konsultationstid er i gennemsnit 20 minutter pr. patient.
Det forudsættes, at betingelserne for at anvende den klassiske køteori er opfyldt.
a) Find sandsynligheden for at venteværelset er fyldt.
b) Find det gennemsnitlige antal patienter i venteværelset.
c) Find den gennemsnitlige ventetid pr. patient.
d) Beregn sandsynligheden for at lægen er optaget.
e) Beregn sandsynligheden for at venteværelset er fuldt besat, såfremt der er 2 læger i systemet med samme gennemsnitlige behandlingstid pr. patient.

Opgave 9.3
I en oliehavn har man 3 oliepumper til losning af lige så mange tankbåde. Såfremt alle pumper er i gang, er der i havneområdet plads til yderligere 2 ventende tankbåde. Såfremt systemet er fuldt besat med tankbåde (i alt 5 både) så henvises tankbådene til nærmeste oliehavn.
Man har erfaring for, at i gennemsnit ankommer der 4 tankbåde pr. døgn, og lossetiden er i gennemsnit 0.5 døgn pr. tankbåd.
Det forudsættes, at betingelserne for at anvende den klassiske køteori er opfyldt.
a) Bestem sandsynligheden for at dette system er blokeret.
b) Find, hvor mange oliepumper der i gennemsnit er i drift.
c) Hvor længe vil de skibe, der ikke omdirigeres til andre havne i gennemsnit opholde sig i havnen?

Opgave 9.4
I forbindelse med en international militær operation skal en signalofficer vurdere om to radiobaserede linkforbindelser kan give en sandsynlighed på 90% for at en samtale straks bliver ekspederet gennem kommunikationssystemet.
Det er ikke muligt at have samtaler i kø, dvs. samtaler der ikke kan ekspederes straks , bliver afvist.
Der skal dimensioneres for 30 samtaleopkald pr. time og hvert opkald varer i middel 1 minut
Anvend den klassiske køteori til at vurdere om kravet til straksekspedition kan opfyldes.

Opgave 9.5
På et militært lager forsøger man kun at have ét eksemplar af en meget dyr enhed på lager. Det sker ved, at straks enheden er afsat så bestilles en ny enhed.
Man har erfaring for, at der i gennemsnit kommer en "kunde" til enheden pr. 4 uger. Såfremt enheden ikke er på lager forsvinder kunden (afvises af systemet)
Det forudsættes, at betingelserne for at anvende den klassiske køteori er opfyldt.
a) Hvad er sandsynligheden for, at enheden ikke er på lager, når der i gennemsnit forløber 1 uge, fra enheden er afsat til den atter er på lager.
b) Besvar samme spørgsmål , hvis der går 2 uger.

Opgave 9.6
Et reservedelsdepot skal reducere omkostningerne. Efter en reorganisering vil der således kun være én ekspedient. Der ankommer i middel en kunde til depotet hvert tiende minut. En kunde kan i middel ekspederes på 5 minutter. Kunder, der ankommer og ikke umiddelbart kan ekspederes, vil alle vente i kø, hvor der er plads nok.
Det forudsættes, at betingelserne for at anvende den klassiske køteori er opfyldt.
a) Bestem trafikudbuddet.
b) Bestem sandsynligheden for straksekspedition.
Depotet forventes udbygget, så der kan stå 5 kunder indendørs.
c) Vurder om kundelokalet er stort nok til at rumme middelantallet af kunder i køsystemet.

Opgave 9.7
Et uniformsdepot skal reorganiseres således at der kun vil være én ekspedient. Der ankommer i middel en kunde til depotet hvert tredie minut. En kunde kan i middel ekspederes på 2 minutter. Kunder, der ankommer og ikke umiddelbart kan ekspederes, vil alle vente i kø, hvor der er plads nok.
Det forudsættes, at betingelserne for at anvende den klassiske køteori er opfyldt.
a) Bestem sandsynligheden for straksekspedition.
Depotet skal udbygges, således at middelantallet af kunder i kø kan vente indendørs og yderligere kunder i kø må vente udenfor.
b) Hvor mange kunder skal depotet dimensioneres til at rumme indendørs?

Opgave 9.8
I forbindelse med en større militærøvelse er det nødvendigt at en bestemt enhed råder over et velfungerende kommunikationssystem.
Styrkechefen har i middel behov for at sende 40 signaler pr. time og transmissionen af et signal varer i middel 1 minut.
Der kan sendes via 2 signallinier og signaler, der ikke umiddelbart kan afsendes, afvises og vil ikke være relevant at sende senere.
Styrkechefens krav til signalsystemet er, at der højst må afvises 1 signal pr. time.
Kan styrkechefens krav til signalsystemet opfyldes?
Det forudsættes, at betingelserne for at anvende den klassiske køteori er opfyldt.

Opgave 9.9
I et apotek er der 4 kasser som hver betjenes af en ekspedienter. Kunderne trækker numre, så den der har det mindste nummer vælger den første ledige ekspedient.
Det viser sig, at der i gennemsnit ankommer 12 kunder pr. minut. Ekspeditionstiderne for hver ekspedient er 15 sekunder pr. kunde.
Det forudsættes, at betingelserne for at anvende den klassiske køteori er opfyldt.
a) Beregn den gennemsnitlige kølængde og det gennemsnitlige antal optagne ekspedienter.
b) Find middelventetiden for de kunder der forsinkes.

Opgave 9.10
I forbindelse med en større militærøvelse er det nødvendigt at en bestemt enhed råder over et velfungerende kommunikationssystem.
Chefen for enheden finder, at der i middel er behov for at sende 30 signaler pr. time og at transmissionen af et signal i middel varer 1 minut.. Der kan kommunikeres via to signalforbindelser. Alle signaler, der ikke straks afsendes vil blive lagt i kø med henblik på afsendelse snarest muligt. Chefens krav til kommunikationssystemet er, at der i middel ikke må være mere end 4 signaler i kø.
Det forudsættes, at betingelserne for at anvende den klassiske køteori er opfyldt.
Beregn den gennemsnitlige kølængde med henblik på, at vurdere om chefens krav til kommunitationssystemet kan opfyldes.

Opgave 9.11
I et militært cafeteria har man kun en kasse, som betjenes af en butiksassistent. Man har observeret, at der i gennemsnit kommer 5 kunder pr. 10 minutter til kassen. Desuden viser det sig, at butiksassistenten i gennemsnit kan betjene 8 kunder på 10 minutter.
Det forudsættes, at betingelserne for at anvende den klassiske køteori er opfyldt.
a) Beregn den gennemsnitlige kølængde ved kassen.
b) Beregn sandsynligheden for at der befinder sig netop 2 kunder i køen.
c) I hvor stor del af tiden er ekspedienten beskæftiget med ekspedition?
d) Hvor længe vil en kunde i gennemsnit befinde sig i køen?
e) Hvor længe vil en kunde i gennemsnit befinde sig i systemet?

10. Hypotesetest

10.1. Indledning

Ofte vil man i statistiske beretninger se vendinger som denne[1]:
"Ved rekrutteringen af officerer til forsvaret er virkningen af TV-spot på antal henvendelser gået frem fra 49% til 52%. Forskellen er lige knap signifikant, hvorimod fremgangen for hjemmesiden er signifikant"

Sådanne statistiske problemer, hvor man ønsker "med 95% sikkerhed, at give et "statistisk bevis" for en eller anden påstand kaldes hypotesetest, og vil blive behandlet i dette kapitel.

10.2. Binomialtest

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

Eksempel 10.1. Binomialtest- 1 variabel.
Ved valget i 2007 stemte 25.5% af vælgerne på Socialdemokraterne. I en opinionsundersøgelse 4 måneder efter valget svarede 1035 vælgere på spørgsmålet om hvilket parti det var mest sandsynligt de ville stemme på, hvis der var valg i morgen. 22.7% svarede, at de ville stemme på partiet Socialdemokraterne. Er dette en signifikant tilbagegang for Socialdemokraterne? ◆

1) **Hvad menes med signifikant?**
   Hvis vi siger, vi statistisk har "bevist" , at socialdemokraterne er gået tilbage, så skal sandsynligheden α for at vi tager fejl naturligvis helst være meget ringe. Sandsynligheden α for at vi tager fejl kaldes **"signifikansniveauet"** .
   Det er derfor strengt taget ikke nok, at sige, at der er en signifikant tilbagegang, man bør samtidig angive på hvilket signifikansniveu
   Indenfor den fagkreds spørgsmålet drejer sig om, er det nok underforstået (ofte 5% eller 10%), men det kan andre jo ikke vide. I det følgende sættes α til 5%.
2) **Forklaring af test "tankegangen".**
   Lad X = antal vælgere der i undersøgelsen vil stemme på Socialdemokraterne .
   Da man enten stemmer på Socialdemokraterne eller på noget andet, så er X binomialfordelt $b(n,p)$
   Ved beregningerne har man som udgangspunkt, at intet er sket siden sidste valg, dvs. vi antager, at der stadig er 25.5 % det stemmer på Socialdemokraterne.
   Dette udtrykkes ved at sige, at **nulhypotesen** $H_0: p = 0.255$ (nul ændring)
   Nulhypotesen skal indeholde en konkret påstand (her et lighedstegn). Den må således aldrig indeholde tegn som $\neq, <$ eller $>$.
   Det man vil bevise placeres (så vidt mulig) i den **alternative hypotese** H
   Her ønsker vi at bevise, at p < 25.5%, så H: p < 0.255

[1] Fra "Analyse af forsvarets elevkampagne 2006 side 8

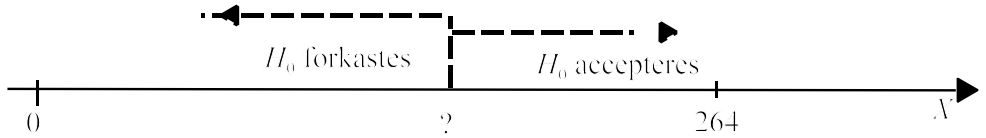
Hvis vi på basis af de følgende beregninger finder ud af, at tilbagegangen for socialdemokraterne er signifikant, så udtrykkes det ofte kort ved at sige, at H_0 **forkastes.**

Vi ved så, at sandsynligheden for at begå fejl ved at påstå dette er mindre end 5%.

Hvis vi omvendt finder, at tilbagegangen ikke er signifikant, så udtrykkes det kort ved at sige, at H_0 accepteres. Bemærk, at vi ikke siger, at der ingen tilbagegang er for socialdemokraterne, for det ved vi faktisk ikke. Vi kan bare ikke bevise det.

Da vi antager, at nulhypotesen er sand, så antages at n = 1035 og p = 0.255.

Da 25.5% af 1035 er 264, og 22.7% af 1035 er ca. 235 kan situationen anskueligøres ved følgende tallinie.



Problemet er om de 235 ligger i forkastelsesområdet, eller acceptområdet.

Hvis grænsen mellem acceptområdet og forkastelsesområdet kaldes x_0 , så er grænsen bestemt ved, at $P(X \le x_0) = \alpha$

Hvis H_0 er sand så er der mindre end 5% chance for at de 235 falder i forkastelsesområdet.

Hvis det så alligevel sker, så må det være fordi H_0 **ikke** er sand. Det betyder at alternativet er sandt, dvs. tilbagegangen er signifikant.

I stedet for at finde x_0 , er det lettere (og også mere oplysende) at beregne $P(X \le 235)$

Hvis denne værdi er mindre end 5% må 235 ligge i forkastelsesområdet.

$P(X \le 235)$ kaldes ***P* - værdien**

3) Opstillling og beregninger

De foregående betragtninger kan sammenfattes på følgende måde:

X = antal vælgere der i undersøgelsen vil stemme på Socialdemokraterne .

X er binomialfordelt $b(n,p)$, hvor n = 1035

Nulhypotese H_0: p = 0.255 Alternativ hypotese H: p < 0.255

Signifikansniveau α = 5%.

Stikprøveværdi :22,7% af 1035 $\approx$ 235

P - værdi = $P(X \le 235)$ =BINOMIALFORDELING(235;1035;0,255;1) = 0,020356.

Denne sandsynlighed er så lille, at den ikke kan skyldes tilfældig variation.

Da P - værdi = 2.04% < 5% forkastes H_0 , dvs. man kan på et signifikansniveau på 5% sige, at socialdemokraterne er i tilbagegang

Resultatet kan sammenfattes i følgende tabel:

Oversigt 10.1 Test af parameter *p* for binomialfordelt variabel

X er binomialfordelt variabel , hvor n er kendt og p er ukendt. Der foreligger en stikprøve på X . Observeret stikprøveværdi x. Signifikansniveau er α . Y er binomialfordelt variabel $b(n, p_0)$, hvor p_0 er en given konstant.		
Alternativ hypotese	Beregning af P - værdi	H_0 forkastes
$H: p > p_0$	P-værdi = $P(Y \geq x)$ =1 - BINOMIALFORDELING(x, n, p_0, 1)	P-værdi $< \alpha$
$H: p < p_0$	P-værdi = $P(Y \leq x)$ = BINOMIALFORDELING(x, n, p_0, 1)	
$H: p = p_0$	P-værdi= $P(Y \geq x)$ for $x > n \cdot p_0$ P-værdi= $P(Y \leq x)$ for $x \leq n \cdot p_0$	P-værdi $< \frac{1}{2}\alpha$

Eksempel 10.2. Opinionsundersøgelse.

Ved valget i 2007 stemte 26.3% af vælgerne på partiet Venstre. I en opinionsundersøgelse 4 måneder efter valget svarede 1035 vælgere på spørgsmålet om hvilket parti det var mest sandsynligt de ville stemme på, hvis der var valg i morgen. 27.2% svarede, at de ville stemme på Venstre. Er dette en signifikant fremgang for Venstre? ◆

Løsning:

X = antal vælgere der i undersøgelsen vil stemme på Venstre .

X er binomialfordelt $b(n,p)$, hvor n = 1035

Nulhypotese H_0: p = 0.263 Alternativ hypotese H: p > 0.263

Signifikansniveau α = 5%.

Stikprøveværdi :27.2% af 1035 ≈ 282

P - værdi = $P(X \geq 282) = 1 - P(X \leq 282)$ =1 - BINOMIALFORDELING(282;1035;0,272;1) = 0,232859

Da P - værdi = 23% > 5% accepteres H_0, dvs. man kan ikke på dette grundlag med 95% sikkerhed sige, at Venstre er i fremgang ◆

Sammenligning af to binomialfordelte variable

I eksempel 10.1 sammenlignede vi en p-værdi for en binomialfordelt variabel med en fast værdi. Ofte skal man som det fremgår af følgende eksempel sammenligne to p-værdier fra to binomialfordelte variable på basis af 2 stikprøver.
Formlerne er ret uoverskuelige.
De er samlet i følgende tabel, og vist hvorledes de bruges i efterfølgende eksempel, men det sikreste er nok eksempelvis at benytte Excel til beregningerne som vist i eksempel 10.3

OVERSIGT 10.2. Test af parametre p_1 og p_2 for binomialfordelte variable.

X_1 og X_2 er **binomialfordelt** henholdsvis $b(n_1, p_1)$ og $b(n_2, p_2)$, hvor n_1 og n_2 er kendte og p_1 og p_2 ukendte.
Observerede stikprøveværdier x_1 og x_2. Signifikansniveau er α

.Lad $\hat{p}_1 = \frac{x_1}{n_1}$, $\hat{p}_2 = \frac{x_2}{n_2}$, $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$ og $u = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p} \cdot (1 - \hat{p}) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$.

Forudsætning:	H: Alternativ hypotese	Beregning	H_0 for-kastes
$5 \le n_1 \cdot \hat{p} \le n_1 - 5$ $5 \le n_2 \cdot \hat{p} \le n_2 - 5$	$H: p_1 > p_2$	P-værdi =1 - NORMALFORDELING(u,0,1,1)	P-værdi $< \alpha$
	$H: p_1 < p_2$	P-værdi = NORMALFORDELING(u,0,1,1)	
	$H: p_1 \neq p_2$	Hvis $\hat{p}_1 \ge \hat{p}_2$ så P-værdi =1 -NORMALFORDELING(u,0,1,1) Hvis $\hat{p}_1 < \hat{p}_2$ så P-værdi =NORMALFORDELING(u,0,1,1)	P-værdi $< \frac{1}{2}\alpha$

Eksempel 10.3. Sammenligning af to binomialfordelte variable

Som nævnt i indledningen påstod man i rapporten fra forsvarets rekruteringstjeneste, at "TV-spot er gået frem fra 49% til 52%. Forskellen er lige knap signifikant".
For at efterprøve denne påstand må man vide antallet i stikprøverne. I 2006 blev det oplyst, at der var 604 svar, mens det kun fremgår af rapporten, at der var færre der svarede i 2005.
Lad os antage, at der var 500 der svarede i 2005.

Løsning:

Vi ønsker at benytte det excel-program, der ligger i www.larsen-net.dk hvor den alternative hypotese er $H: p_1 > p_2$

Da vi skal vise, at der er fremgang fra 2005 til 2006 vælges X_1 og X_2 som nedenfor.

X_1 = antal personer der i 2005 svarer, at TV-spot var der hvor de fik kendskab til uddannelsen

X_2 = antal personer der i 2006 svarer, at TV-spot var der hvor de fik kendskab til uddannelsen

X_1 er binomialfordelt med n = 500 og $p = p_1$. X_2 er binomialfordelt med n = 500 og $p = p_2$

Vi har nu (som ønsket) at den alternative hypotese er $H: p_1 > p_2$.

	A	B	C	D	E	F	G	H
1	**Eksempel 10.3**							
2	X1 = antal personer der svarede i 2005			x1 er binomialfordelt med n1=			500	p=p1
3	X2 = antal personer der svarede i 2006			x2 er binomialfordelt med n2=			604	p=p2
4								
5		**H0: p1=p2 H: p1<p2**						
6								
7	^p1=	0,49	x1=	^p1*n1=	245			
8	^p2=	0,52	x2=	^p2*n2	314,08			
9	^p=	(e7+e8)/(g2+g3)	0,506413					
10								
11	Da ^p*n1=	253,2065217	ligger mellem 5 og		495			
12	og ^p*n2=	305,8734783	ligger mellem 5 og		599			
13	kan approksimeres med normalfordelingen.							
14								
15	y=^p1-^p2=	- 0,03						
16	s=	(C9*(1-C9)*(1/g2+1/g3))^0,5=		0,030228				
17								
18	P-værdi=	NORMFORDELING(b15;0;d14;SAND)=				**0,160491**		
19								
20	Da P-værdi > 0,05 accepteres H0, dvs. **fremgangen i TV-spot er ikke signifikant**							

Beregning ved anvendelse af oversigt 10.2

$x_1 = 0.52 \cdot 604 \approx 314$, $x_2 = 0.49 \cdot 500 \approx 245$

$\hat{p}_1 = 0.52$, $\hat{p}_2 = 0.49$ og $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{314 + 245}{604 + 500} = \frac{559}{1104} = 0.506$.

Da $n_1 \cdot \hat{p} = 604 \cdot 0.506 = 306 \in [5;604-5]$ og $n_2 \cdot \hat{p} = 500 \cdot 0.506 = 253 \in [5;500-5]$
er forudsætningerne for at approksimere med normalfordelingen opfyldt.
Vi finder

$x = \hat{p}_1 - \hat{p}_2 = 0{,}52 - 0{,}49 = 0{,}03$ $s = \sqrt{\hat{p} \cdot (1-\hat{p}) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{0.506 \cdot (1-0.506) \cdot \left(\frac{1}{604} + \frac{1}{500}\right)} = 0.0303$

Da $P(X>0.03)$ =1-NORMFORDELING(0,03;0;0,0303;SAND) = 0.1676 > 0.05 accepteres nulhypotesen. Man kan altså ikke drage en klar konklusion. Fremgangen kunne skyldes udslag af tilfældig variation.

Konklusion: På signifikansniveau 5% er fremgangen i TV - spot ikke signifikant

10.3 Normalfordelingstest

10.3.1. Indledning

Er observationerne sædvanlige målinger, kan man som nævnt i kapitel 2 sædvanligvis antage, at målingerne er normalfordelt. Endvidere vides, at hvis blot antallet i stikprøven er tilstrækkelig stort (over 30) så vil gennemsnittet i alle tilfælde være rimelig normalfordelt.
Vi kan derfor i sådanne tilfælde foretage de følgende omtalte normalfordelingstest.
Testtankegangen er den samme som allerede er gennemgået i afsnit 10.2

10.3.2. Normalfordelingstest (1 variabel)

Kendes spredningen ikke eksakt, kan man anvende en t-test (se oversigt 10.3 og eksempel 10.6)
Kendes spredningen eksakt, kan man anvende en U - test (se oversigt 10.4)

Oversigt 10.3 Test af middelværdi μ for normalfordelt variabel (σ ukendt.)

Der foreligger en stikprøve på X af størrelsen n med gennemsnit $\bar{x}$ og spredning s.
T er en stokastisk variabel der er t - fordelt med $f = n - 1$.

Forudsætninger	Alternativ hypotese	Beregning af P - værdi	H_0 forkastes
σ **ukendt.** Signifikansniveau: α. μ_0 er en given konstant $t = \frac{(\bar{x} - \mu_0)\cdot\sqrt{n}}{s}$.	$H: \mu > \mu_0$	TFORDELING($\lvert t \rvert$, f , 1)	P - værdi $< \alpha$
	$H: \mu < \mu_0$		
	$H: \mu \neq \mu_0$		P - værdi $< \frac{1}{2}\alpha$

Oversigt 10.4 Test af middelværdi μ for normalfordelt variabel (σ kendt eksakt.)

Der foreligger en stikprøve på X af størrelsen n med gennemsnit $\bar{x}$ og eksakt spredning σ
Y er en stokastisk variabel der er normalfordelt med middelværdi μ_0 og spredning $\frac{\sigma}{\sqrt{n}}$

Alternativ hypotese	Beregning af P - værdi	H_0 forkastes
$H: \mu > \mu_0$	$P(Y \geq \bar{x}) = 1$ - NORMALFORDELING($\bar{x}; \mu_0; \sigma/\sqrt{n}; 1$)[2]	P - værdi $< \alpha$
$H: \mu < \mu_0$	$P(Y \leq \bar{x})$ = NORMALFORDELING($\bar{x}; \mu_0; \sigma/\sqrt{n}; 1$)	
$H: \mu \neq \mu_0$	P - værdi = $P(Y \geq \bar{x})$ for $\bar{x} > \mu_0$ P - værdi= $P(Y \leq \bar{x})$ for $\bar{x} \leq \mu_0$	P - værdi $< \frac{1}{2}\alpha$

[2] Er tallene placeret i en søjle med navn "data" kan NORMALFORDELING($\bar{x}, \sigma/\sqrt{n}, 1$) erstattes af 1-ZTEST(data; μ_0 ; σ)

Eksempel 10.4. Normalfordelingstest (1 variabel)

En læge måler til session højden af 20 tilfældigt udvalgte værnepligtige. Resultatet fremgår af nedenstående skema (i cm)

172	185	179	165	192	180	189	170	178	201	165	186	182	191	185	174	178	181	183	179

Middelhøjden var 177.2 på det tidspunkt hvor lægen selv var værnepligtig.
Lægen vil nu gerne på baggrund af ovennævnte materiale, på et signifikansniveau på 5%, vise, at middelhøjden er blevet forøget.
Opstil passende hypoteser og foretag en relevant test.

Løsning:

Lad X = højden af en værnepligtig
X antages normalfordelt med middelværdi μ og spredning σ , der begge ikke kendes eksakt.
Da lægen ønsker at vise, at højden er blevet forøget, bliver
nulhypotesen $H_0: \mu = 177.6$ mod $H: \mu > 177.6$, dvs. vi har en ensidet test.
Excel: (kan findes som en excel-fil på www.larsen-net.dk)
I ark 1 er data placeret og givet navnet "højde". I ark 2 står følgende :

	A	B	C	D
1	**Eksempel 10.4**			
2				
3	X = højden af en værnepligtig	X er normalfordelt med middelværdi μ		
4	H0: μ =	177,2	H:μ >177,2	
5				
6	n =		20	
7	Gennemsnit =	MIDDEL(højde)	180,75	
8	Spredning s =	STDAFV(højde)	8,990492	
9				
10	t =	(C7-B4)*C6^0,5/C8	1,765875	
11	P - værdi =	TFORDELING(ABS(C10);C6-1;1)	0,046741	
12				
13	**Konklusion:** Da P -værdi < 0.05 forkastes H0, dvs.			
14	**vi har vist, at middelhøjden er steget**			

Beregning ved anvendelse af oversigt 10.3
Data indtastes i A1 til A25
Vi finder $\bar{x}$ =MIDDEL(A1:A25)= 180.75 cm og s =STDAFV(A1:A25)= 8.99 cm.

$$t = \frac{(\bar{x} - \mu_0)\cdot\sqrt{n}}{s} = \frac{(180.75 - 177.2)\sqrt{25}}{9.990} = 1.77 . \quad f = 19$$

Da P-værdien = $P(T \geq 1.77)$ = TFORDELING(1,77;19;1) = 0,046741 < 0.05 forkastes H_0 ,
d.v.s. middelhøjden er vokset signifikant ◆

10.3.3 Sammenligning af 2 normalfordelte variable.

Indledning

Ved sammenligning af 2 normalfordelte variable er der afhængigt af hvordan stikprøven er indsamlet valg mellem 2 metoder.
Er stikprøverne for de to variable indsamlet" uafhængigt af hinanden " benyttes sædvanligvis den i oversigt 10.5 angivne metode.
Kendes en tabel over data har Excel et færdigt program som på basis af indtastede data umiddelbart regner relevante P - værdier ud. Dette er vist i eksempel 10.5.
Er de oprindelige data ikke kendt, men kun deres gennemsnit, spredning osv. må man anvende de i oversigt 10.5 angivne formler. Dette er vist i eksempel 10.6

Er observationerne indsamlet "parvist" skal man benytte den i eksempel 10.7 angivne metode.

Er visse forudsætninger opfyldt, kan man dog også anvende den i oversigt 10.6 angivne metode. Denne har en større styrke[1], men man skal altså til gengæld være ret sikker på at forudsætningerne er opfyldt. Et excelprogram er vist i eksempel 10.8

Generel metode

Denne metode kræver, at stikprøverne for de to variable er indsamlet uafhængigt af hinanden. Den kan godt anvendes selv om der er mindre afvigelser fra forudsætningen om at de variable er normalfordelte.
Formlerne fremgår af følgende oversigt

Oversigt 10.5 Test af middelværdier μ_1 og μ_2 og konfidensinterval for differens $\mu_1 - \mu_2$ for 2 normalfordelte variable

Givet 2 stikprøver af X_1 og X_2 med størrelse, gennemsnit og spredning henholdsvis n_1, $\bar{x}_1$, s_1 og n_2, $\bar{x}_2$, s_2.
Signifikansniveau er α. Lad d være en given konstant.

$$a = \frac{s_1^2}{n_1}, b = \frac{s_2^2}{n_2}, \; c = a + b, \; t = \frac{\bar{x}_1 - \bar{x}_2 - d}{\sqrt{c}}$$

Frihedsgradstallet f er det nærmeste hele tal, som er større end $g = \dfrac{c^2}{\dfrac{a^2}{n_1 - 1} + \dfrac{b^2}{n_2 - 1}}$

T er en statistisk variabel der er t - fordelt med frihedsgradstallet f.

Alternativ hypotese	P - værdi	H_0 forkastes
$H: \mu_1 > \mu_2 + d$	TFORDELING($\lvert t \rvert, f, 1$)	P - værdi $< \alpha$
$H: \mu_1 < \mu_2 + d$		
$H: \mu_1 \neq \mu_2 + d$		P - værdi $< \frac{1}{2}\alpha$

95% Konfidensinterval for differens $\mu_1 - \mu_2$: $\bar{x}_1 - \bar{x}_2 - t_{0,975}(f) \cdot \sqrt{c} \le \mu_1 - \mu_2 \le \bar{x}_1 - \bar{x}_2 + t_{0,975}(f) \cdot \sqrt{c}$

[1] En test har "større styrke" end en anden, hvis den med givne data og et givet signifikansniveau giver den største chance for at forkaste nulhypotesen.

Eksempel 10.5. Sammenligning af 2 normalfordelte variable (data givet)

To produktionsmetoder M1 og M2 ønskes sammenlignet. Der udvælges tilfældigt 20 personer, hvoraf de 10 bliver sat til at arbejde med den ene metode, og de 10 andre med den anden. To personer bliver syge under forsøgsperoden, så der var kun 8 personer i den ene gruppe.
Efter 2 ugers forløb, beregnede man for hver person det gennemsnitlige tidsforbrug pr enhed. Da metode 1 er mere kostbar end metode 2, ønsker man kun at gå over til den, hvis tidsforbruget pr enhed ved metode 1 er mindst 2 minutter mindre end ved metode 2.
Man fik følgende resultater.

M_1	87.8	91.9	89.8	89.0	92.6	89.4	91.4	88.7		
M_2	92.4	94.6	93.0	94.0	92.4	92.9	96.4	92.1	92.8	91,4

1) Undersøg på basis af disse resultater, om det på et signifikansniveau på 5% kan påvises at tidsforbruget ved metode M_1 er 2 minutter mindre end ved metode M_2
2) Angiv endvidere et 95% konfidensinterval for differensen mellem de to middeludbytter.

Løsning:

1) Lad X_1 = udbyttet ved anvendelse af metode M_1 og
X_2 = udbyttet ved anvendelse af metode M_2.
X_1 og X_2 antages approksimativt normalfordelte med middelværdi og spredning henholdsvis μ_1, σ_1 og μ_2, σ_2.

$H_0: \mu_2 = \mu_1 + 2$ $\quad$ $H: \mu_2 < \mu_1 + 2$ eller

$H_0: \mu_2 - \mu_1 = 2$ $\quad$ $H: \mu_2 - \mu_1 < 2$

Bemærk: Hypoteserne skal skrives så differensen bliver positiv, da Excel ikke tillader negative forskelle.
Vi vælger den robuste metode, som ikke forudsætter eksempelvis at varianserne er ens
Excel kalder metoden "to stikprøver med forskellig varians"
Excel løsning:(kan findes som en excel-fil på www.larsen-net.dk)
Lad stikprøveværdierne for M_1 og M_2 stå i to søjler i cellerne A7 til A14 og B71 til B16

Funktioner ▶ Dataanalyse ▶ t-test: to stikprøver med forskellig varians ▶ Den fremkomne tabel udfyldes ▶ "Område for variabel 1": B7..B16 ▶ "Område for variabel 2" A7 ..A14 ▶ Afmærk "Etiketter" ▶ "Hypotese for forskel i middelværdi": 2
(**bemærk**, at da Excel ikke tillader en negativ forskel, må vi sætte M_2 søjlen ind som variabel 1 og M_1 søjlen ind som variabel 2.)

	A		B	C	D	E	F
1	**Eksempel 10.5**						
2	X1= tidsforbrug ved anvendelse af metode M1				X1 har middelværdi μ1 og spredning σ1		
3	X2= tidsforbrug ved anvendelse af metode M2				X2 har middelværdi μ2 og spredning σ2		
4	H0: μ2 =μ1 +2	H:μ2 < μ1 +2					
5	**Data:**						
6	M1	M2		**Resultat**			
7	87,8	92,4		t-test: To stikprøver med forskellig varians			
8	91,9	94,6					
9	89,8	93			*M2*	*M1*	
10	89	94		Middelværdi	93,2	90,075	
11	92,6	92,4		Varians	2,095556	2,887857	
12	89,4	92,9		Observationer	10	8	
13	91,4	96,4		Hypotese for forskel i middelværdi	2		
14	88,7	92,1		fg	14		
15		92,8		t-stat	1,489397		
16		91,4		P(T<=t) en-halet	0,079282		P-værdi
17				t-kritisk en-halet	1,761309		
18				P(T<=t) to-halet	0,158563		
19				t-kritisk to-halet	2,144789		
20	**1) Konklusion**						
21	Da P-værdi =	0,0792815	> 0,05 accepteres H0, dvs.				
22	**vi kan ikke på dette grundlag bevise at tidsforbruget ved metode 1 er 2 minutter kortere end ved metode 2**						
23	**2) Konfidensinterval**						
24			r =	E19*(E11/E12+F11/F12)^0,5=	1,620043		
25		**Nedre grænse =**		E10-F10-E24	**1,504957**		
26		**Øvre grænse =**		E10-F10+E24	**4,745043**		

Bemærk, at da testen er ensidet sammenlignes "en-halet" P - værdien med 0.05.
Havde den været tosidet skulle vi have sammenlignet den "to-halet P - værdi på 0.158 med 0.05.[1]
Bemærk, at vi sammenligner altid P - værdien med signifikansniveauet α
Konklusion: Vi kan ikke på dette grundlag bevise, at tidsforbruget ved metode 1 er 2 minutter mindre end ved metode 2 (vi er dog tæt på en forkastelse).

2) **95% Konfidensinterval for differens**
Formlen der benyttes er

$$\mu_1 - \mu_2 : \bar{x}_1 - \bar{x}_2 - t_{0,975}(f)\cdot\sqrt{c} \le \mu_1 - \mu_2 \le \bar{x}_1 - \bar{x}_2 + t_{0,975}(f)\cdot\sqrt{c}\ ,\ \text{hvor}\ c = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$$

Excel kan ikke umiddelbart beregne konfidensintervallet, men da alle tallene kan findes i cellerne, er det, som det ses af excelløsningen, ret let at foretage beregningerne.
95% konfidensinterval: $1.50 \le \mu_2 - \mu_1 \le 4.75$ ♦

[1] t-kritisk en-halet, er den grænse, som t-stat - værdien skal over, for at man kan forkaste nulhypotesen, dvs $t_{0.95}(14) = 1.7613$
t-kritisk to-halet, er tilsvarende $t_{0.975}(14) = 2.1447$

Eksempel 10.6. Sammenligning af 2 normalfordelte variable (oprindelige data ikke givet)

Et luftfartsselskab A hævder, at dets fly til USA i gennemsnit afgår mere præcist end et konkurrerende luftfartsselskab.
En forbrugergruppe undersøger denne påstand ved i en given periode at bestemme forsinkelserne for samtlige flyafgange til USA for hver af de to selskaber.
Man fandt følgende tal:

Luftfartsselskab	Antal afgange	$\bar{x}$	s
A	100	55 minutter	30 minutter
B	80	60 minutter	35 minutter

Støtter undersøgelsen luftfartsselskab A's påstand?

Løsning:

X_A = forsinkelsen i minutter for luftfartselskab A.
X_B =forsinkelsen i minutter for luftfartselskab B.
X_A og X_B antages approkssimativt normalfordelte med middelværdi og spredning henholdsvis μ_A, σ_A og μ_B, σ_B.
Da vi ønsker at vise, at A er mere præcise end B, så haves:

$H_0: \mu_A = \mu_B$ $\quad$ $H: \mu_A < \mu_B$

	A	B	C	D	E
1	**Eksempel 10.6**				
2					
3	XA =forsinkelsen for luftfartselskab A			XA er normalfordelt med middelværdi μA	
4	XB =forsinkelsen for luftfartselskab A			XB er normalfordelt med middelværdi μB	
5	H0: μA =μB	H: μA < μB			
6	**Data**		**Beregning**		
7	nA =	100	a=	B9^2/B7	9
8	x-streg-A=	55	b=	B12^2/B10	15,3125
9	sA =	30	c=	E7+E8	24,3125
10	nB =	80	t=	(B8-B11-B13)/KVROD(E9)	-1,01404
11	x-streg-B=	60	g=	E9^2/(E7^2/(B7-1)+E8^2/(B10-1))	156,1194
12	sB =	35	f =	RUND.OP(E11;0)	157
13	d=	0	P-værdi=	TFORDELING(ABS(E10);E12;1)	0,156062
14		**Konklusion:** Da p -værdi > 0.05 accepteres H0, dvs.			
15		**det kan ikke på dette grundlag vises, at A er mere præcis end B**			
16					

Parvise observationer

Parvise observationer (Matched pairs samples) anvendes, hvis det er sådan, at når man har valgt et objekt i den ene gruppe, så er det samtidig givet hvem der skal være i den anden gruppe (der er **ikke** uafhængighed).
Eksempel: Hvis man vil undersøge om der er en sammenhæng mellem ægtefællers intelligenskvotient (IQ), så er det selvfølgelig således, at er en person er valgt (randomiseret) , så vil vedkommendes ægtefælle automatisk også blive valgt.
Lad os igen betragte problemstillingen i eksempel 10.5, men nu antage, at forsøget er foretaget på en anden måde.

Eksempel 10.7. Parvise observationer

To produktionsmetoder M1 og M2 ønskes sammenlignet. Der udvælges tilfældigt 8 personer. Efter lodtrækning bliver 4 personer sat til først i 2 uger, at arbejde med produktionsmetode M1 og derefter i de næste 2 uger med produktionsmetode M2.
De øvrige 4 personer arbejder omvendt først med metode M2 og derefter med metode M1.
Efter 2 ugers forløb, beregnede man for hver person det gennemsnitlige tidsforbrug pr. enhed.
Da metode 1 er mere kostbar end metode 2, ønsker man kun at gå over til den, hvis tidsforbruget pr. enhed ved metode 1 er mindst 2 minutter mindre end ved metode 2.
Man fik følgende resultater.

Person nr.	1	2	3	4	5	6	7	8
M_1	87.8	91.9	89.8	89.0	92.6	89.4	91.4	88.7
M_2	92.4	94.6	93.0	94.0	92.4	92.9	96.4	92.1

1) Undersøg på basis af disse resultater, om det på et signifikansniveau på 5% kan påvises at tidsforbruget ved metode M_1 er 2 minutter mindre end ved metode M_2
2) Angiv endvidere et 95% konfidensinterval for differensen mellem de to middeludbytter.

Forklaring på metode:
Da en forsøgsperson kan være hurtig og en anden langsom (person 1 er således hurtigere end person 2) kan spredningen på M_1 og M_2 være så stor, at man intet kan vise.
Hvis man i stedet tager differenserne M2 - M1 vil disse forskelle jo udjævnes, da person 1 jo er hurtig under arbejdet med begge metoder, mens person 2 er langsom ved begge.

Person nr.	1	2	3	4	5	6	7	8
M_1	87.8	91.9	89.8	89.0	92.6	89.4	91.4	88.7
M_2	92.4	94.6	93.0	94.0	92.4	92.9	96.4	92.1
$D = M_2 - M_1$	4.6	2.7	3.2	5	-0.2	3.5	5	3.4

I stedet for at benytte metoden i eksempel 10.5 kan vi nu teste nulhypotesen
$H_0: D = 0$ mod $H: D < 2$ ved metoden i eksempel 10.4 (en variabel)
Dette sker automatisk i Excel.

Løsning:
1) Lad X_1 = udbyttet ved anvendelse af metode M_1 og
X_2 = udbyttet ved anvendelse af metode M_2.
X_1 og X_2 antages approksimativt normalfordelte med middelværdi og spredning henholdsvis μ_1, σ_1 og μ_2, σ_2.
$H_0: \mu_2 = \mu_1 + 2$ $H: \mu_2 < \mu_1 + 2$ eller

$H_0: \mu_2 - \mu_1 = 2$ $\quad\quad$ $H: \mu_2 - \mu_1 < 2$

Data indtastes (de samme som i eksempel 10.5) og derefter:
Funktioner ▶ Dataanalyse ▶ t-test: Parvis dobbelt stikprøve for middelværdi ▶ Den fremkomne tabel udfyldes ▶ "Område for variabel 1": B7:B14 ▶ "Område for variabel 2" A7:A14 ▶ Afmærk "Etiketter" ▶ "Hypotese for forskel i middelværdi": 2

Den følgende excel-udskrift kan findes som en excel-fil på www.larsen-net.dk .

	A	B	C	D	E	F
1	**Eksempel 10.7**					
2	X1= tidsforbrug ved anvendelse af metode M1				X1 har middelværdi μ1 og spredning σ1	
3	X2= tidsforbrug ved anvendelse af metode M2				X2 har middelværdi μ2 og spredning σ2	
4	H0: μ2 =μ1 +2 H:μ2 < μ1 +2					
5	**Data:**					
6	M1	M2		**Resultat**		
7	87,8	92,4		t-test: Parvis dobbelt stikprøve for middelværdi		
8	91,9	94,6				
9	89,8	93			M2	M1
10	89	94		Middelværdi	93,475	90,075
11	92,6	92,4		Varians	2,122142857	2,887857143
12	89,4	92,9		Observationer	8	8
13	91,4	96,4		Pearson-korrelation	0,433089341	
14	88,7	92,1		Hypotese for forskel i middelværdi	2	
15				fg	7	
16				t-stat	2,339141989	
17				P(T<=t) en-halet	0,025955245	P-værdi
18				t-kritisk en-halet	1,894578604	
19				P(T<=t) to-halet	0,05191049	
20	**1) Konklusion**			t-kritisk to-halet	2,364624251	
21	Da P-værdi =	0,0259552	< 0,05 forkastes H0, dvs			
22		**det er bevist, at tidsforbruget ved metode 1 er 2 minutter kortere end ved metode 2**				
23	**2) Konfidensinterval**					
24		gennemsnit	d=	MIDDEL(A26:A33)	3,4	
25	Differens			*Differens*		
26	4,6					
27	2,7	r =		Konfidensniveau(95,0%)	1,41525139	
28	3,2	**nedre grænse**		E24-E27	1,98474861	
29	5	**Øvre grænse**		E24+E27	4,81525139	
30	-0,2					
31	3,5					
32	5					
33	3,4					

Konklusion: M1 er signifikant 2 minutter lavere end M2, dvs. man vil gå over til at benytte metode M1

2) Konfidensinterval for differens: $1.99 \le \mu_2 - \mu_1 \le 4.82$

Metode der kræver forudsætninger opfyldt.

Denne metode kræver,

1) at man har samme antal gentagelser af hver variabel
2) at de to variable har næsten samme varians. Kravet er at man kan få en accept ved en varianstest. Til gengæld har testen så en større styrke[1] end den "robuste test".

Eksempel 10.8. Sammenligning af 2 normalfordelte variable (ens varians)

To produktionsmetoder M1 og M2 ønskes sammenlignet. Der udvælges tilfældigt 16 personer, hvoraf de 10 bliver sat til at arbejde med den ene metode, og de 10 andre med den anden.
Efter 2 ugers forløb, beregnede man for hver person det gennemsnitlige tidsforbrug pr. enhed.
Da metode 1 er mere kostbar end metode 2, ønsker man kun at gå over til den, hvis tidsforbruget pr. enhed ved metode 1 er mindst 2 minutter mindre end ved metode 2.
Man fik følgende resultater.

M_1	87.8	91.9	89.8	89.0	92.6	89.4	91.4	88.7
M_2	92.4	94.6	93.0	94.0	92.4	92.9	96.4	92.1

Man mener, at spredningerne er rimelig ens.

1) Undersøg på basis af disse resultater, om det på et signifikansniveau på 5% kan påvises at tidsforbruget ved metode M_1 er 2 minutter mindre end ved metode M_2
2) Angiv endvidere et 95% konfidensinterval for differensen mellem de to middeludbytter.

Løsning:

1) Først testes om spredningerne er ens: $H_0: \sigma_1^2 = \sigma_2^2$:
   Vælg "Funktioner", "Dataanalyse", "F-test: Dobbelt stikprøve for varians"
   Den fremkomne tabel udfyldes:
   "Område for variabel 1": A1:A8 "Område for variabel 2": B1:B8
   "Outputområde": Skriv cellenummer for øverste venstre celle i det ønskede outputområde.

F-test: Dobbelt stikprøve for varians

	M2	*M1*	
Middelværdi	93,475	90,075	
Varians	2,122143	2,887857	
Observationer	8	8	
fg	7	7	
F	0,73485		
P(F<=f) en-halet	0,34732		H_0 accepteres, da P-værdi=0.347 > 0.025
F-kritisk en-halet	0,264058		

   Da H_0 accepteres, kan vi i det følgende tillade os at fortsætte beregningerne under den forudsætning, at de er ens.

2) Nu testes: $H_0: \mu_2 \le \mu_1 + 2$ mod $H: \mu_2 > \mu_1 + 2$
   Vælg "Funktioner", "Dataanalyse", "t-test: to stikprøver med ens varians"
   Den fremkomne tabel udfyldes:
   "Område for variabel 1": B1:B8 "Område for variabel 2": A1:A8
   "Hypotese for forskel i middelværdi": 2
   (bemærk, at da Excel ikke tillader en negativ forskel, må vi sætte K_2 søjlen ind som variabel 1 og K_1 søjlen ind som variabel 2.
   "Outputområde": Skriv cellenummer for øverste venstre celle i det ønskede outputområde.

[1] En test har "større styrke" end en anden, hvis den med givne data og et givet signifikansniveau giver den største chance for at forkaste nulhypotesen.

t-test: To stikprøver med ens varians

	M2	*M1*	
Middelværdi	93,475	90,075	
Varians	2,122143	2,887857	
Observationer	8	8	
Puljevarians	2,505		
Hypotese for forskel i middelværdi	2		
fg	14		
t-stat	1,769107		
P(T<=t) en-halet	0,049323		H_0 forkastes, da P-værdi = 0.049 < 0.05
t-kritisk en-halet	1,761309		
P(T<=t) to-halet	0,098646		
t-kritisk to-halet	2,144789		

M1 er altså signifikant 2 minutter lavere end M1, dvs. man vil gå over til at benytte metode M1

2) Et 95% konfidensinterval for differensen er (se eventuelt oversigt 10.1) $\bar{x}_2 - \bar{x}_1 \pm t_{0.975}(14) \cdot s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$

$$r = t_{0.975}(14) \cdot s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = (t-\text{kritisk to}-\text{halet}) \cdot \sqrt{\text{puljevarians}} \cdot \sqrt{\frac{1}{\text{observation 1}}+\frac{1}{\text{observation 2}}}$$

Vi finder 95% konfidensinterval: $2.0 \le \mu_2 - \mu_1 \le 4.39$

Nedenstående oversigt angiver de formler, der skal anvendes:

Oversigt 10.6. Test af middelværdier μ_1 og μ_2 og konfidensinterval for differens $\mu_1 - \mu_2$ for 2 normalfordelte variable. $\sigma_1 \approx \sigma_2$

Givet 2 stikprøver af X_1 og X_2 med størrelse, gennemsnit og spredning henholdsvis n_1, $\bar{x}_1$, s_1 og n_2, $\bar{x}_2$, s_2. Signifikansniveau er α. Lad d være en given konstant

$$t = \frac{\bar{x}_1 - \bar{x}_2 - d}{s \cdot \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \text{ hvor } s^2 = \frac{(n_1-1)\cdot s_1^2 + (n_2-1)\cdot s_2^2}{n_1+n_2-2},\ f = n_1+n_2-2.$$

T er en statistisk variabel der er t - fordelt med frihedsgradstallet f.

Forudsætning.	Alternativ hypotese	P - værdi	H_0 forkastes
$H_0: \sigma_1 = \sigma_2$ accepteres ved F - test	$H: \mu_1 > \mu_2 + d$	TFORDELING($\lvert t \rvert$, f, 1)	P - værdi< α
	$H: \mu_1 < \mu_2 + d$		
	$H: \mu_1 \neq \mu_2 + d$		P - værdi < $\frac{1}{2}\alpha$
F - test: $F = \frac{s_1^2}{s_2^2}$ Q er F - fordelt[1] $F(n_1-1, n_2-1)$	$H: \sigma_1^2 \neq \sigma_2^2$	$P(Q \ge F)$ = Ffordeling(F,n1,n2) for F >1 $P(Q \le F)$ = 1 - Ffordeling(F,n1,n2) for F <1	P - værdi < $\frac{1}{2}\alpha$

95% konfidensinterval for differens $\mu_1 - \mu_2$: $\bar{x}_1 - \bar{x}_2 - t_{0.975}(f) \cdot s \cdot \sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \le \mu_1 - \mu_2 \le \bar{x}_1 - \bar{x}_2 + t_{0,975}(f) \cdot s \cdot \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$

1) Excel: F-fordeling: Lad F = 2, n_1=6, n_2=8, $P(Q \ge F)$ = FFORDELING(2;6;8)=0,1792

10.4 Test i Antalstabeller

10.4.1 Indledning

Vi har hidtil betragtet tilfælde, hvor antal variable var 1 eller 2.
Man kan naturligvis godt have mere end 2 variable. Vi vil i dette kapitel betragte observationer, som bliver katalogiseret i klasser, eksempelvis kan klasserne være en opdeling af vælgere efter hvilket politisk parti de vil stemme på.
Eksempel 10.8 illustrerer dette.

Eksempel 10.8 (kategoriske data)

Et ministerium planlægger en oplysningskampagne om de fysiske og psykiske virkninger af at ryge hash. Før kampagnen viste en undersøgelse at 7% af indbyggerne ønskede at hash blev legaliseret, 65% at man beholde det nuværende forholdsvis liberale straffepolitik, 18 % ønskede strengere straffe og 10 % havde ingen mening. Efter kampagnen spurgte man 500 personer (repræsentativt udvalgt) , og svarene fremgik af følgende tabel.

	Legalisering	Efter eksisterende lov	Strengere straf	Ingen mening
Efter kampagnen	39	336	99	26

Kan man på dette grundlag vise på et signifikansniveau på $\alpha = 0.01$,at kampagnen har betydet en ændring af folks mening? ♦

Vi vil se på tilfældige eksperimenter, der opfylder følgende krav:

1. Eksperimentet skal have k mulige udfald, hvor $k \geq 2$.
   Disse udfald kaldes "klasser", " kategorier" eller "celler"
2. Eksperimentet gentages n gange uafhængigt af hinanden.
3. Sandsynligheden for de k udfald er $p_1, p_2, \ldots, p_k$ (hvor $p_1 + p_2 + \ldots + p_k = 1$) er de samme ved de n gentagelser.
4. De statistiske variable der er af interesse er antallet $n_1, n_2, \ldots, n_k$ i hver af de k klasser.

Betragter vi eksempel 10.8 ses, at betingelserne er opfyldt:
Eksperimentet består i tilfældigt at udtage $n = 500$ personer af en stor population og spørge dem om strafferammen for besiddelse af hash
1) Udfaldene er svar på spørgsmålet, og der er $k = 4$ (og kun 4) svarmuligheder (4 klasser).
2) Resultatet af hvad en person svarer vil være uafhængigt af hvad de øvrige svarer.
3) Sandsynligheden for udfaldene i de 4 klasser vil være p_1, p_2, p_3 og p_4 , hvor disse sandsynligheder er ukendte.
4) De statistiske variable X_i er antal personer blandt 500 som har en af de i meninger om straffen for hash.

Test i antalstabeller kaldes sædvanligvis for χ^2 - test (udtales ki-i-anden) efter den testfunktion der anvendes.

10.4.2 En-vejs tabel

Eksempel 10.8.(fortsat)

Kan man på det i eksempel 10.8 angivne grundlag vise på et signifikansniveau på $\alpha = 0.01$,at kampagnen har betydet en ændring af folks mening?

Løsning:

Lad

X_1= antal personer blandt 500 , der går ind for legalisering. $P(X_1) = p_1$.

X_2= antal personer blandt 500 , der går ind for den nuværende straffepolitik. $P(X_2) = p_2$.

X_3= antal personer blandt 500 , der går ind for en strengere straf. $P(X_3) = p_3$.

X_4= antal personer blandt 500 , der ingen mening har. $P(X_4) = p_4$.

Vi ønsker at teste nulhypotesen

$H_0 : p_1 = 0.07,\ p_2 = 0.65,\ p_3 = 0.18,\ p_4 = 0.10$

mod den alternative

H: "Mindst én af sandsynlighederne afviger fra den angivne værdi i nulhypotesen."

Vi beregner nu de forventede værdier, forudsat nulhypotesen er sand.
Resultatet opstilles i det følgende skema:

nr	1	2	3	4
O_i	39	336	99	26
E_i	$E_1 = 500 \cdot 0.07 = 35$	$E_2 = 500 \cdot 0.65 = 325$	$E_3 = 500 \cdot 0.18 = 90$	$E_4 = 500 \cdot 0.1 = 50$

Betingelserne for, at man kan udføre en χ^2 - test er, at ingen af klasserne har en forventet værdi på mindre end 1, og mindst 80% af klasserne skal have en forventet værdi over 5.

Da alle de forventede værdier i dette eksempel er større end 5 er forudsætningerne opfyldt.
Man danner nu summen

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(39-35)^2}{35} + \frac{(336-325)^2}{325} + \frac{(99-90)^2}{90} + \frac{(26-50)^2}{50} = 13.249 .$$

Det er klart, at jo mere de forventede værdier afviger fra de observerede værdier jo større bliver χ^2 . Vi forkaster derfor nulhypotesen ved en ensidet såkaldt χ^2 -test med frihedsgradstallet k - 1 hvor k er antal klasser .

Excel - løsning: (kan findes som en excel-fil på www.larsen-net.dk) .

	A		B	C	D	E	F	G		
1	**Eksempel 10.8**									
2	x1=antal personer der går ind for legalisering P(X1)=p1									
3	x2=antal personer der går ind for nuværende politik P(X2)=p2									
4	x3=antal personer der går ind for strengere straf P(X3)=p3									
5	x4=antal personer der ingen mening har P(X4)=p4									
6										
7	**H0: p1=**	**0,07**	**p2=**	**0,65**	**p3=**	**0,2**	**p4=**	**0,1**		
8	**H: Mindst en af sandsynlighederne afviger fra den angivne værdi i nulhypotesen**									
9	n =	500								
10			x1	x2	x3	x4				
11	P-værdier		0,07	0,65	0,18	0,1				
12	Observerede data		39	336	99	26				
13	Forventede data		35	325	90	50				
14	Beregninger		C11*$B9 og kopiere mod højre							
15										
16	P-værdi =		CHITEST(C12:F12;C13:F13)=					**0,004127**		
17										
18	Da P- værdi < 0.05 forkastes H0, dvs.									
19	**kampagnen har haft en betydning for meningen om legalisering af hash**									

Da P - værdi =0.42% <0.05 (endda mindre end 1%)forkastes nulhypotesen, dvs. kampagnen har haft en betydning for meningen om legalisering af hash. ♦

10.4.3. To-vejs tabel

I dette afsnit betragter vi eksperimenter hvor data er karakteriseret ved to kriterier.
Et eksempel herpå er følgende:

Eksempel 10.11. (testning af uafhængighed)

Ved en uddannelsesinstitution indstillede et år 500 studerende sig til en årsprøve, der bl.a. omfattede matematik og fysik.
De opnåede karakterer i de to fag inddeltes i 4 grupper:

Observerede værdier		Fysikkarakterer				Total
		-3, 0	2	4, 7	10, 12	
Matematik-karakterer	- 3, 0	18	46	13	0	77
	2	22	60	42	5	129
	4, 7	7	123	42	16	188
	10, 12	2	28	68	8	106
Total		49	257	165	29	500

Undersøg om der er en sammenhæng mellem de opnåede fysikkarakterer og de opnåede matematikkarakterer.

Løsning:
X_1 = antal studerende med opnået matematikkarakter
X_2 = antal studerende med opnået fysikkarakterer
H_0: X_1 og X_2 er statistisk uafhængige.

Udfaldene antages at være uafhængige, og de studerende antages at være repræsentative for en årgang med samme sandsynlighed fra årgang til årgang.
Vi får følgende tabel over de forventede værdier:

Forventede værdier		Fysikkarakterer			
		0, 3, 5	6, 7	8, 9	10, 11, 13
Matematik-karakterer	0, 3, 5	$\frac{77 \cdot 49}{500} = 7.546$	$\frac{77 \cdot 257}{500} = 39.578$	$\frac{77 \cdot 165}{500} = 25.41$	$\frac{77 \cdot 29}{500} = 4.466$
	6, 7	$\frac{129 \cdot 49}{500} = 12.642$	$\frac{129 \cdot 257}{500} = 66.306$	$\frac{129 \cdot 165}{500} = 42.57$	$\frac{129 \cdot 29}{500} = 7.482$
	8, 9	$\frac{188 \cdot 49}{500} = 18.424$	$\frac{188 \cdot 257}{500} = 96.632$	$\frac{188 \cdot 165}{500} = 62.04$	$\frac{188 \cdot 29}{500} = 10.904$
	10, 11, 13	$\frac{106 \cdot 49}{500} = 10.388$	$\frac{106 \cdot 257}{500} = 54.484$	$\frac{106 \cdot 165}{500} = 34.98$	$\frac{106 \cdot 29}{500} = 6.148$

Da alle de forventede værdier er over 1, og kun 1 klasse af 16 ligger under 5 er betingelserne for en χ^2 - test opfyldt.
Vi beregner nu teststørrelsen

$$\chi^2 = \frac{(18 - 7.546)^2}{7.546} + \frac{(46 - 39.578)^2}{39.578} + \ldots + \frac{(8 - 6.148)^2}{6.148} = 108.917$$

Det kan vises, at for 2-vejs tabeller er frihedsgradstallet f = (antal søjler - 1)*(antal rækker -1)
Vi har følgelig $f = (4-1)\cdot(4-1) = 9$

Excel - løsning:

	A	B	C	D	E	F	G	
1	**Eksempel 10.9**							
2	X1 = antal studerende med opnået matamatikkarakter							
3	x2 = antal studerende med opnåede fysikkarakter							
4								
5	H0: X1 og X2 er statistisk uafhængige							
6								
7	**Observerede data**							
8				**fysik**				
9			**- 3; 0**	**2**	**4;7**	**10;12**	**sum**	
10		**- 3; 0**	18	46	13	0	77	
11	**Matematik**	**2**	22	60	42	5	129	
12		**4;7**	7	123	42	16	188	
13		**10;12**	2	28	68	8	106	
14		**SUM**	49	257	165	29	500	
15								
16								
17	**Forventede data**			**fysik**				
18			**- 3; 0**	**2**	**4;7**	**10;12**		
19		**- 3; 0**	7,546	39,578	25,41	4,466		
20	**Matematik**	**2**	12,642	66,306	42,57	7,482		
21		**4;7**	18,424	96,632	62,04	10,904		
22		**10;12**	10,388	54,484	34,98	6,148		
23								
24	Beregningseksempel i celle c19				$G10*C$14/G14			
25								
26	P-værdi =	CHITEST(C10:F13;C19:F22) =			**2,44228E-19**			
27								
28	Da P- værdi < 0.05 forkastes H0, dvs. **matematik og fysikkarakterer er ikke uafhængige**.							

Det er ret besværligt at beregne de forventede værdier i Excel, men er det først gjort, er det nemt at få P-værdien.
Bemærk at ønsker man at kopiere ned gennem en søjle, skal visse adresser være absolutte (med $ tegn)

Da P - værdi = $2.44 \cdot 10^{-19} < 0.05$ forkastes nulhypotesen (stærkt) dvs.
der er ikke uafhængighed mellem fysikkaraktererne og matematikkaraktererne.
Når vi ser på tallene er det tydeligt at gode karakterer i det ene fag også har en tendens til at bevirke gode karakterer i det andet fag.

Opgaver

Opgave 10.1
En ny vaccine formodes med en sandsynlighed på mindst 85% at have en forebyggende virkning over for en bestemt influenzatype. Før en truende influenzaepedemi vaccineres et hospitalspersonale på 600 personer med den pågældende vaccine. 125 af disse bliver smittet af sygdommen. Kan dette opfattes som en eksperimentel påvisning af, at vaccinen er mindre virksom end ventet?

Opgave 10.2.
I forbindelse med anskaffelse af en ny uniform til søværnet fik 100 soldater (tilfældigt udvalgt) i en uge udleveret uniformstype A og i en anden uge udleveret uniformstype B. Hvilken uniform man fik først blev valgt ved lodtrækning.
Det viste sig efterfølgende , at af de 100 soldater foretrak 65 soldater uniformstype A.
Kan dette tages som et statistisk bevis for, at soldaterne generelt foretrækker uniformstype A.

Opgave 10.3
Det forventes, at lovgivningen bliver strammet omkring mængden af skadelige partikler i bilers udstødningsgas. En person mener, at mere end 20% af forsvarets biler ikke vil opfylde de forventede nye krav. Ved en undersøgelse af 30 af forsvarets biler tilfældigt udvalgt, fandt man, at 10 af disse ikke kunne opfylde de nye krav.
Test om dette på et signifikansniveau på 5 % er et bevis for, at mere end 20% af forsvarets biler udsender flere skadelige partikler end ønskeligt.

Opgave 10.4
En fabrikant af chip til computere reklamerer med, at højst 2% af en bestemt type chip, som fabrikken sender ud på markedet er defekte.
Et stort computerfirma, vil købe et meget stort parti af disse chip, hvis påstanden er rigtigt. For at teste påstanden købes 1000 af dem. Det viser sig, at 33 ud af de 1000 er defekte.
Kan fabrikantens påstand på denne baggrund forkastes på signifikansniveau 5% ?

Opgave 10.5
Forsvaret overvejer at indføre et nyt sikkerhedssystem. Før man indfører dette mere kostbare system, ønsker man at sikre sig, at det nye system giver væsentlig færre arbejdsulykker.
Man har to arbejdspladser A og B, som er nogenlunde af samme størrelse og i sikkerhedsmæssig henseende er nogenlunde ens. I arbejdsplads B indføres det nye system, og efter en passende forsøgsperiode tælles antallet af (større eller mindre) arbejdsulykker på de to arbejdspladser.
Man fandt, at i B havde 5 ud af 250 medarbejdere været udsat for en arbejdsulykke, mens det for A gjaldt for 24 ud af 263 medarbejdere.
Kan man på et signifikansniveau på 1% (da systemet er kostbart at indføre) vise, at det nye sikkerhedssystem giver færre ulykker.

Opgave 10.6
I en opinionsundersøgelse svarede 893 personer, hvoraf 475 var mænd og 418 kvinder.
Er stikprøven repræsentativ med hensyn til kønsfordeling, idet vi antager, at antallet af mænd og kvinder i hele befolkningen er ens.

Opgave 10.7
Ved en undersøgelse af en eventuel sammenhæng mellem luftforurening og forekomsten af lungecancer sammenlignedes bl.a. sygdommens forekomst i byen X - køb ing inden for den gamle bygrænse (i nærheden af byens industrivirksomheder) med dens forekomst i samme bys forstadsområde (villakvarter):

	Antal tilfælde af lungecancer	Samlet indbyggerantal
Indre by	30	9000
Forstadsområde	40	27000

1) Det ses. at den relative hyppighed af cancertilfælde i den indre by afviger fra den relative hyppighed i forstadsområdet. Kan dette forklares som et tilfældigt udsving? Den opstillede nulhypotese. som testes, ønskes specificeret med angivelse af den alternative hypotese.
2) Diskuter muligheden for at drage årsagsmæssige konklusioner ud fra det fundne testresultat.

Opgave 10.8
Mange forbrugere tror, at såkaldte "mandagsbiler", dvs. biler produceret om mandagen, har flere alvorlige fejl end biler produceret på ugens øvrige arbejdsdage.
For at undersøge, om der er noget grundlag for denne tro, udtog man på en bilfabrik tilfældigt 100 "mandagsbiler" og undersøgte dem for fejl. Man fandt at 8 biler havde alvorlige fejl. Tilsvarende udtog man tilfældigt 200 biler, der var produceret på ugens øvrige arbejdsdage, og man fandt 12 biler, der havde alvorlige fejl.
Giver denne undersøgelse støtte til formodningen om, at "mandagsbiler" er af dårligere kvalitet end andre biler.

Opgave 10.9
I en nordisk undersøgelse om anvendelsen af it i folkeskolen blev der bl.a. spurgt:
Hvor ofte anvende it på skolen til at kommunikere meddelelser ud til alle lærere på skolen .
I Sverige svarede 31 skoler på spørgsmålet, og af dem var der 21, der svarede, at det skete mindst en gang om ugen. I Danmark var det 40 ud af 47 skoler der svarede det samme.
Kan man heraf slutte, at man i de danske skoler er signifikant bedre til at anvende it til den form for kommunikation?

Opgave 10.10

En virksomhed bliver af miljøkontrollen pålagt at formindske indholdet i sit spildevand af et stof A, der mistænkes for at kunne forurene grundvandet. Indholdet af stoffet A i spildevandet skal under 1.7 mg/l, og miljøkontrollen henviser til en ny metode, som burde kunne formindske indholdet til det ønskede niveau. For at vurdere den nye metode ønskes foretaget 15 delforsøg (ét forsøg om dagen i 3 uger).

1) Følgende værdier af indholdet af A fandtes (i mg/l).

1.55	1.66	1.98	1.37	1.60	1.62	1.53	1.79	1.72	1.57	1.70	1.40	1.77	1.58	1.84

   Kan man ud fra disse data bevise på signifikansniveau $\alpha = 0.05$, at indholdet af A ved benyttelse af den ny metode er under 1.7 mg/l.
2) Angiv et estimat for middelindholdet af stoffet A ved den nye metode, og et 95% tosidet konfidensinterval for det forventede indhold af A .

Opgave 10.11

Den samlede mængde kosmisk stråling , som en person blev udsat for ved flyvning med jetfly fra Kastrup til New York kunne på et givet tidspunkt erfaringsmæssigt antages at være normalfordelt med middelværdien $\mu = 5.42$ mrem og spredningen $\sigma = 0.55$ mrem.

1) Beregn sandsynligheden p_2 for, at mængden af kosmisk stråling, som personen udsættes for, mindst er 6.25 mrem.

Nogle år senere foretog man 12 målinger af den kosmiske straling, som en person blev udsat for ved at flyve den samme strækning. Resultaterne var:

4.85	6.24	5.10	6.78	5.09	6.62	4.45	7.09	5.30	5.46	6.14	6.49

Det antages i det følgende spørgsmål fortsat, at mængden af kosmisk straling er normalfordelt men middelværdien og spredningen antages nu at være ukendt.

2) Foretag en testning af , om middelværdien kan antages at være over 5.42.

Opgave 10.12

På pakken af en iscreme står, at portionen indeholder 14 gram fedt. For at kontrollere dette købes 10 pakker is, og fedtindholdet måles. Man finder et gennemsnit på 13.1 gram og et estimat s for spredningen på 0.42 gram.

1) Kan man ud fra disse data bevise på signifikansniveau $\alpha = 0.01$, at middelindholdet afviger fra 14 gram?
2) Angiv et estimat for middelindholdet.
3) Angiv et 95% konfidensinterval for middelindholdet.

Opgave 10.13

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

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

I ubeskadede skove er SO_2 - koncentrationen 20 μ g/m^3. Giver forsøgene et bevis for, at middelkoncentrationen af SO_2 i den beskadigede skov er større end normalt?

Angiv et tosidet 95%-konfidensinterval for SO_2 - koncentrationen.

Opgave 10.14

På et bestemt mejeri produceres mælk i 1 liters kartoner. Idet mælkens massefylde er 1.033, bør vægten være 1033 gram. Det vides, at vægten er normalfordelt med en middelværdi på 1045 gram (for at undgå underfyldte kartoner) og med en spredning på 6.96 gram.

1) Beregn sandsynligheden for at en tilfældig karton mælk vejer mindre end 1033 gram.
2) Ledelsen kræver at mejeriet skal justere processen, dvs. ændre middelværdien μ, således at i middel 1% af produktionen er underfyldt.
   Find μ, idet spredningen antages uændret 6,96 gram.
3) Ved en mejeriinspektion udtages en stikprøve på 20 kartoner mælk. Man vejede disse og fandt et gennemsnit på $\bar{x}$ =1040 gram og estimat for spredningen på s = 7.66 gram.
   Kan man ud fra disse data bevise, på et signifikansniveau på $\alpha = 0.001$, at middelvægten er over 1033 gram.

Opgave 10.15

Et levnedsmiddelfirma havde udviklet en diæt, som har lavt indhold af fedt, kulhydrater og kolesterol. Diæten er udviklet med henblik på patienter med hjerteproblemer, men firmaet ønsker nu at undersøge diætens virkning på folk med vægtproblemer.

To stikprøver på hver 100 personer med vægtproblemer blev udtaget tilfældigt. Gruppe A fik den nye diæt, mens gruppe B fik den diæt, man normalt gav. For hver person blev registreret størrelsen af vægttabet i en 3 ugers periode.

Man fandt følgende værdier for gennemsnit og spredning:

Gruppe A: $\bar{x}_A = 9.31$ kg, $s_A = 4.67$

Gruppe B: $\bar{x}_B = 7.40$ kg, $s_B = 4.04$.

1) Undersøg om vægttabet for gruppe A er signifikant større end for gruppe B. Signifikansniveau $\alpha = 5\%$.
2) Beregn et 95% konfidensinterval for differensen mellem de to gruppers middelværdier.

Opgave 10.16

Det påstås at modstanden i en tråd af type A er større end modstanden i en tråd af type B. Til afklaring af denne påstand udtages tilfældigt 6 tråde af hver type og deres modstande måles. Følgende resultater fandtes:

Modstand i tråd A (i ohm)	0.140	0.138	0.143	0.142	0.144	0.137
Modstand i tråd B (i ohm)	0.135	0.140	0.142	0.136	0.138	0.140

Hvilke konklusioner kan drages med hensyn til påstanden?

Opgave 10.17

I forsvaret ønsker man ved et forsøg at undersøge om et skift fra en dæktype A til en anden dæktype B ville formindske benzinforbruget væsentligt.
Blandt 24 "ens" biler blev (randomiseret) valgt 12 biler som blev forsynet med dæk af type A, og de resterende 12 biler blev forsynet med dæk af type B
De blev derefter sat til at køre en bestemt rute og deres benzinforbrug blev målt.
Resultaterne var (i kilometer pr. liter)

A	10.5	10.6	11.4	11.8	12.6	10.8	11.0	11.6	12.8	10.4	12.0	10.8
B	11.3	11.4	12.3	12.1	12.9	11.4	11.5	12.1	13.9	10.9	12.7	11.5

Viser disse resultater på et signifikansniveau på 5%, at dæktype B giver et mindre benzinforbrug end type A? Bemærk: Jo mindre benzinforbrug, jo flere kilometer kan der køres pr. liter benzin.

Opgave 10.18

En produktion af plastikvarer må omlægges på grund af bestemmelser i en ny miljølov.
Ved den fremtidige produktion kan inden for miljølovens rammer vælges mellem 2 produktionsmetoder I og II. Metode I er den dyreste, og fabrikanten har regnet ud, at det (kun) kan betale sig at benytte metode I, såfremt den giver et middeludbytte, som er mindst 10 måleenheder (udbytteprocenter) større end udbyttet ved benyttelse af metode II.
1)Ved et fuldstændigt randomiseret forsøg fandtes følgende måleresultater:

Metode I	35.2	38.1	37.6	37.6	34.9	37.9	36.5	40.0	36.2	37.4	37.2	37.9
Metode II	26.2	22.2	24.3	24.5	22.0	27.6	23.8	22.8	23.4	20.8		

Fabrikanten valgte herefter at benytte metode I.
Foretag en undersøgelse af, om valget var statistisk velmotiveret.
Opstil et 95% - konfidensinterval for differensen mellem middeludbytterne ved benyttelse af metoderne l og II.

Opgave 10.19

Måling af intelligenskvotient på 16 tilfældigt udvalgte studerende ved en studie (med mere end 200 studerende) viste et gennemsnit på $\bar{x}_1 = 107$ og en varians på $s_1^2 = 100$, medens en tilsvarende måling på 14 tilfældigt udvalgte studerende fra en anden studieretning viste et gennemsnit på $\bar{x}_2 = 112$ og en varians på $s_2^2 = 64$.
Tyder disse tal på en forskel på studentermaterialet på de to retninger?

Opgave 10.20

l) 100 studerende, 52 piger og 48 drenge, indstillede sig til en prøve, ved hvilken 39 piger og 27 drenge bestod. Undersøg. om det anførte tyder på, at resultatet ved den pågældende prøve afhænger af deltagerens køn.
2) Det oplyses supplerende, at pigerne ved ovennævnte prøve opnåede et gennemsnit på 64% med en empirisk spredning på 10%, medens drengenes gennemsnit var 59% med en empirisk spredning på 8%. Undersøg, om det anførte kan tages som vidnesbyrd om, at piger i almindelighed klarer sig bedre end drenge ved den omhandlede prøve.

Opgave 10.21

I forbindelse med den fysiske træning af nye soldater blev der klaget over, at træningsprogrammet var for hårdt. Som begrundelse blev det bl. a. påstået, at soldaterne i middel tabte sig mere end 2 kg i løbet af træningsperioden. For at undersøge denne påstand blev udtaget 16 tilfældigt valgte rekrutter og deres vægt blev målt før og efter træningsprogrammet.

Resultaterne (målt i kg) før og efter træningsprogrammet blev for hver person følgende:

nr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Før	89.5	74.0	69.5	91.0	75.0	81.0	64.0	75.0	70.0	61.5	76.5	67.5	85.5	74.5	106.5	57.0
Efter	82	69.5	70	87.5	72.5	79.0	63.5	74.0	66.0	60	73.5	65.0	83.5	70.5	99.5	54

Foretag en testning af, om soldaterne taber sig mere end 2 kg.

Opgave 10.22

Et bestemt medikament ønskes testet for dets effekt på blodtrykket. 12 mænd fik deres blodtryk målt før og efter indtagelse af medikamentet. Resultaterne var:

mand nr	1	2	3	4	5	6	7	8	9	10	11	12
Før	120	124	130	118	140	128	140	135	126	130	126	127
Efter	128	131	131	127	132	125	141	137	118	132	129	135

Udfør en testning af, om disse tal tyder på, at medikamentet påvirker blodtrykket.

Opgave 10.23

Et diætprodukt påstår i en reklame, at brug af produktet i en måned vil resultere i et vægttab på 3 kg.

En forbrugerorganisation ønsker at teste denne påstand. 8 personer bruger produktet i en måned, og det resultat fremgår af nedenstående tabel:

person nr	1	2	3	4	5	6	7	8
Startvægt	81	101	98	99	78	71	75	93
Slutvægt	78	95	95	97	73	69	70	89

1) Undersøg på grundlag af disse tal, om det på basis af disse tal på et signifikansniveau på 5% kan vises, at reklamens påstand er fejlagtig?
2) Opstil et 95% tosidet konfidensinterval for middelværdien af vægttabet, og giv på grundlag heraf en vurdering af virkningen af diætproduktet.

Opgave 10.24

En producent af malervarer har laboratorieresultater, der tyder på, at en ny lak A, har en større slidstyrke end den sædvanlige lak B. Han ønsker en afprøvning i praksis og aftaler med ejerne af 6 bygninger med mange trapper, at han må lakere deres trapper. Efter 3 måneders forløb måles graden af slid (i %) i hver bygning.

1) Angiv hvorledes du ville foretage forsøget.

2) De målte værdier af slid efter valg af plan var

Bygning nr	1	2	3	4	5	6
Ny lak	20.3	25.1	21.8	19.6	18.9	23.5
Sædvanlig lak	19.5	28.4	21.6	22.0	20.9	25.8

Undersøg om observationerne leverer et eksperimentelt bevis for, at den nye lak er mere slidstærk end den sædvanlige lak.

Opgave 10.25

Følgende tabel angiver i procent valgresultatet ved sidste folketingsvalg, og ved en meningsmåling, hvor 1035 vælgere bliv spurgt om hvem de ville stemme på hvis der var valg i morgen

Partier	A	B	C	F	K	O	V	Ø	Øvrige
Valg 2001 i %	25.9	9.2	10.3	6	1.7	13.2	29	3.4	1.3
Opinion 2005	24.8	10.6	9.7	8.2	1.4	13.8	27.2	3.6	0.7

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

Kan man på dette grundlag vise på et signifikansniveau på 5%, at der er sket en ændring af partiernes tilslutning.

Opgave 10.26

I faget statistik ønskede man at undersøge om 3 hold var lige dygtige. Ved en prøve blev hver besvarelse karakteriseret som enten tilfredsstillende eller ikke-tilfredsstillende.

Man fik følgende resultat:

Klasse	x	y	z
Holdstørrelse	31	25	36
Antal tilfredsstillende besvarelser	22	16	26

Kan man på dette grundlag vise på et signifikansniveau på 5%, vise, at der er forskel på klassernes statistikkundskaber.

Opgave 10.27
En børnelæge vil gerne undersøge om der antallet af børn der bliver født er ligelig fordelt på alle ugens dage. Der udvælges tilfældigt 500 personer, og man optæller hvad ugedag de er født på. Resultatet var:

Ugedag	søndag	mandag	tirsdag	onsdag	torsdag	fredag	lørdag
Antal	57	78	74	76	71	81	63

Test på et signifikansniveau på 5% om antallet er ligeligt fordelt på dagene.

Opgave 10.28
En terning kastedes 120 gange, hvorved følgende resu1tater fandtes:

	Antal Øjne					
	1	2	3	4	5	6
Antal gange	25	17	15	23	24	16

Test nulhypotesen: Terningen er en ærlig" terning.

Opgave 10.29
En kemika1iefabrik har påbegyndt en fabrikation af kunstgødning. Ved fabrikationen hældes gødningen i sække af 5 "ens" maskiner, idet det tilstræbes, at nettoindho1det i sækkene er 25 kg i hver .
Ved indkørselen af produktionen fandt man, at der var mange overvægtige og undervægtige sække. Fø1gende antalstabel indeholder produktionsresultatet ved første prøvekørsel:

		Maskiner				
		1	2	3	4	5
Nettovægt	Under 24 kg	5	3	7	3	12
	Mellem 24 og 26 kg	14	17	16	15	13
	Over 26 kg	11	10	7	12	5

Foretag en testning af, om det kan antages, at vægtfordelingen er den samme for de 5 maskiner.

Opgave 10.30
Ved start af en stor amerikansk industrivirksomhed underkastedes alle 173 ansøgere til et bestemt job på fabrikken en psykoteknisk prøve. Idet ansøgerne grupperedes efter, om de var medlemmer af en fagforening eller ikke, er nedenstående anført resultaterne af den pågældende prøve.

	Resultat af prøven		
	godt	middel	dårligt
Medlem af en fagforening	37	42	23
Ikke medlem af en fagforening	17	26	28

Hvad kan der sluttes om sammenhæng mellem præstation ved prøven og medlemskab af en fagforening?

Opgave 10.31

En fabrik, der arbejdede i 3 - holdskift, fremstillede b1.a. en bestemt maskindel i massefabrikation.

For at undersøge, om kvaliteten af denne maskindel påvirkedes af omstændigheder, der afhang af, inden for hvi1ket tidsrum af døgnet fabrikationen fandt sted (træthed, belysningsforhold m.v.), lod man et bestemt arbejdshold arbejde på hvert af de 3 skift en uge ad gangen. Man regnede med, at produktionsbetingelserne fra uge ti1 uge var i det væsentlige uændrede.

Arbejdsholdets ugentlige produktion var:

Skift	Antal ikke - defekte emner	Antal defekte emner
kl. 8^{00} - 16^{00}	1602	88
kl. 16^{00} - 24^{00}	1590	122
kl. 0^{00} - 8^{00}	1507	103

Foretag en statistisk ana1yse af, om produktionens kvalitet må antages at afhænge af produktionsperioden.

11. Tidsrækker

11.1. Indledning.

Erhvervsmæssige beslutninger baseres ofte på prognoser. Disse er naturligvis usikre, da de baserer sig på en uændret udvikling. De kan derfor ikke opfange en pludselig ændring såsom politiske beslutninger, oliepriserne himmelflugt osv.
Grundideen i prognoser er en opbyning i tidsserier. Sædvanligvis kan man opdele eksempelvis en virksomheds drift over en periode i

1) en "trend" (en langsigtet tendens) ,
2) en konjukturmæssigt mønster (svingninger omkring trenden i perioder af et antal år)
3) et sæsonmæssigt mønster (udsving indenfor et år, lavsæson, højsæson)
4) støj , dvs, uforklarlige, tilfældige udsving, som ikke følger et bestemt mønster.

En analyse bygger på, at man starter med at undersøge om der er et sæsonmæssigt mønster, og hvis det er tilfældet, så prøve at eliminere det, så trenden bliver tydeliger.
Hvis talmaterialet rækker tilstrækkelig langt i tid kan man så tilsvarende forsøge at finde et konjukturmæssigt mønster. Tilsidst prøver man så at få opstillet et udtryk for trenden, og ud fra den sige noget om fremtiden, med den usikkerhed som støjen som bl.a. støjen bevirker.

11.2. Det sæsonmæssige mønster og korrektion heraf.

11.2.1 Indledning

Korrektionen foregår i en række trin.

1) Først undersøges om der overhovedet er sæsonmæssige udsving ved at tegne en graf.
2) Beregning af glidende gennemsnit
3) Beregning af sæsonfaktorer
4) Foretage en ændring af de aktuelle tal (ved benyttelse af sæsonfaktorerne) til tal der er "renset" for sæsonudsving (f. eks skal man måske sætte de faktiske tal 10% op for at korrigere for sæsonvariationen)

Vi anvender følgende gennemgående eksempel for en virksomheds kvartalsvise omsætning. Tallene er på forhånd korrigeret for prisudviklingen, dvs. der er tale om et mængdeindeks.[1]
Eksemplet findes på adressen i www.larsen-net.dk

[1] Man vælger et "sammenligningsår" f.eks 2000 , hvor talstørrelsen sættes lig med 100. Indekstallene i de følgende år fremkommer som årets talværdi i procent af tallet i 2000. Mængdeindeks $= \frac{\text{indeks for omsætning}}{\text{prisindeks}}$

Eksempel 11.1. Beregning af sæsonfaktorer m.m.

Den følgende tabel viser en virksomheds kvartalsmæssige mængdeindeks i perioden 2002 - 2005.

Tabel 11.1: Kvartalsvis mængdeindeks for virksomhed A' s omsætning. Indeks 2001 = 100

år	2002	2003	2004	2005
1 kvartal	94	96	108	117
2 kvartal	103	108	113	124
3 kvartal	95	100	106	121
4 kvartal	117	120	132	142

11.2.2 Grafisk undersøgelse

En figur tegnes i Excel:

Vælg på værktøjslinien "Guiden Diagram ▶ kurve ▶ marker ønsket figur ▶ Næste ▶ marker udskriftsområde ▶ Næste ▶ Næste ▶ Udfør

Cursor på Y-akse. højre musetast ▶ Formater akse ▶ skala ▶ minimum = 90 ▶ ok

Cursor på X-akse. højre musetast ▶ Formater akse ▶ "flueben" væk ved Værdiakse(Y) krydser mellem kategoreserierværdier ▶ ok

Kvartal	Indeks
02-k1	94
02-k2	103
02-k3	95
02-k4	117
03-k1	96
03-k2	108
03-k3	100
03-k4	120
04-k1	108
04-k2	113
04-k3	106
04-k4	132
05-k1	117
05-k2	124
05-k3	121
05-k4	142

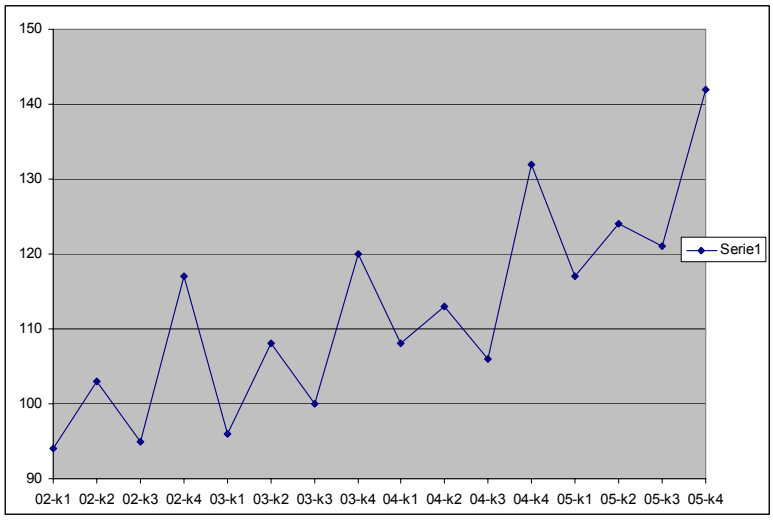


Fig 11.1. *Mængdeindeks i 16 kvartaler*

Vi ser, at der er en stigende tendens i tallene, samt en nogenlunde regelmæssig sæsonbevægelse Der er en tydelig højsæson i 4 kvartal (julehandel) samt en vis lavsæson i 1. og 3. kvartal.

11.2.3. Beregning af centreret glidende gennemsnitstal

Et centreret glidende gennemsnit beregnes på følgende måde:

For tredie kvartal i år 2002: $C_{02}^{3} = \frac{\frac{1}{2}\cdot 94 + 103 + 95 + 117 + \frac{1}{2}\cdot 96}{4} = 102.5$

Fjerde kvartal (rykker et kvartal frem): $C_{02}^{4} = \frac{\frac{1}{2}\cdot 103 + 95 + 117 + 96 + \frac{1}{2}\cdot 108}{4} = 103.375$ osv.

I Excel skrives formlen i den første celle =(B1/2+B2+B3+B4+B5/2)/4 og kopieres derefter.

Kvartal	Indeks	C
02-k1	94	
02-k2	103	
02-k3	95	102,5
02-k4	117	103,375
03-k1	96	104,625
03-k2	108	105,625
03-k3	100	107,5
03-k4	120	109,625
04-k1	108	111
04-k2	113	113,25
04-k3	106	115,875
04-k4	132	118,375
05-k1	117	121,625
05-k2	124	124,75
05-k3	121	
05-k4	142	

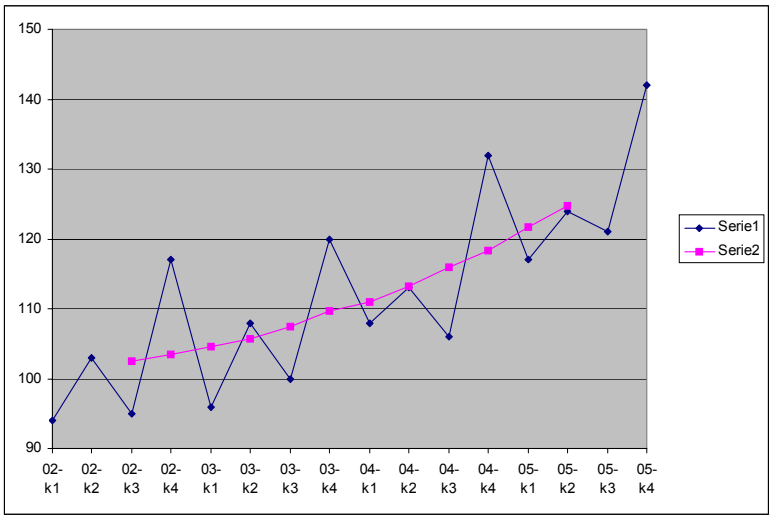


Fig 11.2: *Centreret glidende indeks*

Bemærk, at man kun kan bruge centreret glidende gennemsnit omkring et nogenlunde retlinet udviklingsforløb. I tilfælde af en meget krum kurve må man eventuelt først foretage en passende transformation (se næste kapitel)

11.2.4 Beregning af sæsonfaktorer ved multiplikativ og additiv model

Indledning:

En sæsonfaktor for en given periode (et kvartal i eksemplet) udtrykker i hvilken retning og hvor meget det målte tal skal korrigeres (f. eks. 5% op).

Hvorledes det skal beregnes afhænger af, om sæsonsvingningerne er større, jo højere absolut værdi der er tale om (en multiplikativ model), eller er af samme absolutte størrelse uanset niveauet (en additiv model)

Betragtes figur 11.2, synes udsvingene at have nogenlunde samme størrelse uanset niveauet, hvilket kunne tale for en additiv model. Imidlertid må man forvente, at for en industrivirksomheds produktion vil udsvingene blive større jo højere omsætningen er.

Beregning af sæsonfaktorer ved multiplikativ model

1) For hvert kvartal i hvert af årene beregnes $A = \frac{Y}{C}$, hvor Y er det målte tal, og C er det glidende gennemsnitstal. Er A > 1 er der "højsæson" og er A < 1 er der "lavsæson".

   I 1 kvartal fås således (se Excel data) $\frac{96}{104.625} = 0.9176$, $\frac{108}{111} = 0.973$ og $\frac{117}{121.625} = 0.962$

Tilsvarende beregnes for 2. kvartal osv.

2) Sæsonfaktoren for 1. kvartal beregnes nu som gennemsnittet af alle A-tallene i 1. kvartal, dvs.

$$\frac{0.918+0.973+0.962}{3}=0.951$$

Tilsvarende beregnes sæsonfaktorerne for 2, 3 og 4. kvartal.

I Excel fås

kvartal	Y	C	A	sæsonfaktorer	korrektion
02-k1	94				
02-k2	103				
02-k3	95	102,5	0,926829	0,923946894	0,925477158
02-k4	117	103,375	1,131802	1,11384761	1,115692392
03-k1	96	104,625	0,917563	0,950836325	0,952411124
03-k2	108	105,625	1,022485	1,004755226	1,006419326
03-k3	100	107,5	0,930233		
03-k4	120	109,625	1,094641		
04-k1	108	111	0,972973		
04-k2	113	113,25	0,997792		
04-k3	106	115,875	0,914779		
04-k4	132	118,375	1,1151		
05-k1	117	121,625	0,961973		
05-k2	124	124,75	0,993988		
05-k3	121				
05-k4	142				
sum				3,993386056	4

Den sidste søjle "korrektion" skyldes, at da sæsonfaktorernes sum burde være 4, men er 3.9933, så korrigerer man ved, at multiplicerer hver sæsonfaktor med 4/ 3.9933.

At sæsonfaktorernes sum ikke er 4 skyldes, at de yderste led jo ikke er regnet med.

Det er vigtigt, at der bør indgå et rimeligt stort antal år i beregningerne og i hvert fald mindst 4.

Endvidere må man være opmærksom på ekstreme talværdier i en tidsserie. De kan eksempelvis skyldes en arbejdskonflikt, og må i sådant et tilfælde naturligvis udelades.

Sæsonkorrektion af tidsserie

For en multiplikativ model beregnes et sæsonkorrigeret tal Z for en given periode (kvartal) af formlen $Z=\dfrac{Y}{\text{sæsonfaktor(korrigeret)}}$.

Eksempelvis for kvartal 03-k1: $Z=\dfrac{96}{0.9524}=100.8$

Beregningerne sker i Excel:

kvartal	Y	C	A	sæsonfaktorer	korrektion	Z
02-k1	94					98,69687329
02-k2	103					102,3430267
02-k3	95	102,5	0,926829	0,923946894	0,925477158	102,6497512
02-k4	117	103,375	1,131802	1,11384761	1,115692392	104,8676148
03-k1	96	104,625	0,917563	0,950836325	0,952411124	100,7968068
03-k2	108	105,625	1,022485	1,004755226	1,006419326	107,3111348
03-k3	100	107,5	0,930233			108,0523697
03-k4	120	109,625	1,094641			107,556528
04-k1	108	111	0,972973			113,3964076
04-k2	113	113,25	0,997792			112,2792429
04-k3	106	115,875	0,914779			114,5355119
04-k4	132	118,375	1,1151			118,3121808
05-k1	117	121,625	0,961973			122,8461082
05-k2	124	124,75	0,993988			123,2090807
05-k3	121					130,7433673
05-k4	142					127,2752248
sum				3,993386056	4	

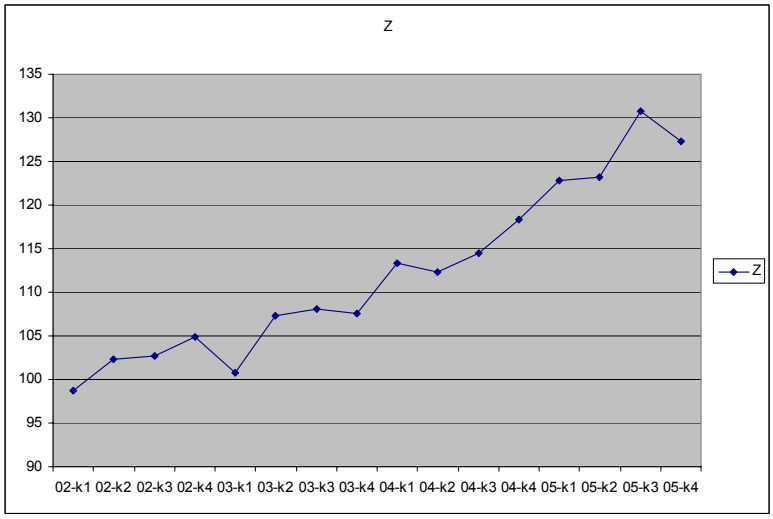


Fig 11.3. *Sæsonkorrigeret tidsserie*

Figuren er tegnet ved at placere kvartal og Z søjlen ved siden af hinanden

Sammenlignes figur 11.1 med figur 11,3 ses, at de store sæsonudsving er dæmpet kraftigt ned. Idet vi ikke her vil behandle egentlige konjukturudsving (som kræver tal fra væsentlig flere år end de 4), så vil næste trin være at lave en prognose for fremtiden ved benyttelse af regressionsanalyse Dette sker i næste kapitel.

Beregning af sæsonfaktorer ved additiv model

Er sæsonsvingningerne af samme absolutte størrelse uanset niveauet bør man anvende en additiv model.

1) For hvert kvartal i hvert af årene beregnes $A = Y - C$, hvor Y er det målte tal, og C er det glidende gennemsnitstal. Er A > 0 er der "højsæson" og er A < 0 er der "lavsæson".
   I 1 kvartal fås således (se Excel data) $96 - 104.625 = -8.625$, $108 - 111 = -3$ og $117 - 121.625 = -4.625$.
   Tilsvarende beregnes for 2. kvartal osv.
2) Sæsonfaktoren for 1. kvartal beregnes nu som gennemsnittet af alle A-tallene i 1. kvartal, dvs.
   $$-\frac{8.625 + 3 + 4.625}{3} = -5.417$$
   Tilsvarende beregnes sæsonfaktorerne for 2, 3 og 4. kvartal.

Kvartal	Y	C	A	sæsonfaktorer	korrektion
02-k1	94				
02-k2	103				
02-k3	95	102,5	-7,5	-8,29167	-8,11458
02-k4	117	103,375	13,625	12,54167	12,71875
03-k1	96	104,625	-8,625	-5,41667	-5,23958
03-k2	108	105,625	2,375	0,458333	0,635417
03-k3	100	107,5	-7,5		
03-k4	120	109,625	10,375		
04-k1	108	111	-3		
04-k2	113	113,25	-0,25		
04-k3	106	115,875	-9,875		
04-k4	132	118,375	13,625		
05-k1	117	121,625	-4,625		
05-k2	124	124,75	-0,75		
05-k3	121				
05-k4	142				
sum				-0,70833	0

   Den sidste søjle "korrektion" skyldes, at da sæsonfaktorernes sum burde være 0, men er -0.70833, så korrigerer man ved, at subtrahere hver sæsonfaktor med -0.70833/4.

Sæsonkorrektion af tidsserie

Det sæsonkorrigerede tal Z for en given periode (kvartal) beregnes af formlen
Z = Y - korrigeret sæsonfaktor
Eksempelvis for kvartal 03-k1: $Z = 96 - (-5.23958) = 101.24$

Beregningerne i Excel:

Kvartal	Y	C	A	sæsonfaktorer	korrektion	Z
02-k1	94					99,23958
02-k2	103					102,3646
02-k3	95	102,5	-7,5	-8,29167	-8,11458	103,1146
02-k4	117	103,375	13,625	12,54167	12,71875	104,2813
03-k1	96	104,625	-8,625	-5,41667	-5,23958	101,2396
03-k2	108	105,625	2,375	0,458333	0,635417	107,3646
03-k3	100	107,5	-7,5			108,1146
03-k4	120	109,625	10,375			107,2813
04-k1	108	111	-3			113,2396
04-k2	113	113,25	-0,25			112,3646
04-k3	106	115,875	-9,875			114,1146
04-k4	132	118,375	13,625			119,2813
05-k1	117	121,625	-4,625			122,2396
05-k2	124	124,75	-0,75			123,3646
05-k3	121					129,1146
05-k4	142					129,2813
sum				-0,70833	0	

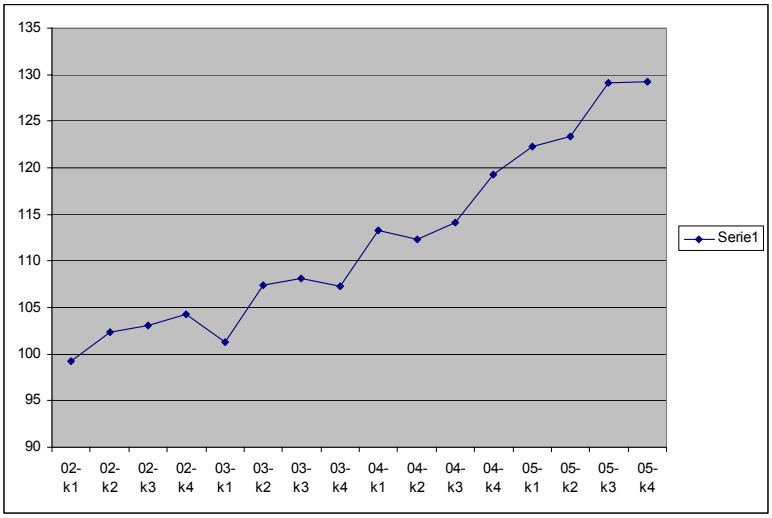


Fig. 11.4 *Sæsonkorrigeret tidsserie*

Sammenlignes figur 11.1 med figur 11.4 ses, at de store sæsonudsving er dæmpet kraftigt ned.

Opgaver

Opgave 11.1
En virksomhed W's kvartalsvise omsætning målt i mængder fremgår af nedenstående skema .

Tabel 11.1: Kvartalsvis mængdeindeks for virksomhed A' s omsætning. Indeks 2001 = 100				
år	2002	2003	2004	2005
1 kvartal	103	108	116	132
2 kvartal	112	123	132	150
3 kvartal	105	115	123	139
4 kvartal	115	127	141	154

1) Foretag en grafisk undersøgelse af kvartalstallene
2) Beregn sæsonfaktorerne for kvartalerne
3) Foretag en sæsonkorrektion af omsætningstallene i tabellen.
4) Foretag en sæsonkorektion for den efterfølgende "aktuelle" 1. kvartal 2006, hvor det faktiske tal er 141.
5) Tegn det sæsonkorrigerede mængdeindeks for omsætningen i hele perioden fra 1 kvartal 2002 til første kvartal 2006. Den aktuelle tendens i virksomhedens omsætning kommenteres.

Opgave 11.2
På adressen http://www.statistikbanken.dk/pris6
"fra industriens ordre og omsætningssituation (ikke sæsonkorrigeret) "
afmærk "ordreindgang fra hjemmemarkedet", tobaksindustri, og måned = 2000 til 2004
1) Flyt indekstallene over i Excel (skriv dem kolonnevis)
2) Saml dem i kvartaler.
3) Foretag en grafisk undersøgelse af kvartalstallene
4) Foretag en sæsonkorrektion af kvartalstallene
5) Foretag en sæsonkorrektion for den efterfølgende "aktuelle" 1. kvartal 2005" idet man fra statistikbanken kan beregne det aktuelle kvartalstal.
6) Tegn det sæsonkorrigerede indekstal for hele perioden fra 1. kvartal 2000 til 1. kvartal 2005.

Opgave 11.3
På adressen http://www.statistikbanken.dk under " Miljø og energi" og derefter "energi" findes nogle oplysninger om Danmarks forbrug af energi efter type og mængde.
1) Hent forbrug af "olie, i alt" ind målt i tons for perioden 2000 til 2005 (i måneder) ind i Excel.
2) I stedet for at opgøre forbruget månedsvis, så opgør det i kvartaler, og tegn og kommenter udviklingen af forbruget.
3) Beregn et centreret glidende gennemsnit af forbrugstallene i tabellen og tegn denne.
Benyt i det følgende en additiv model.
4) Beregn sæsonfaktorer, og foretag en sæsonkorrektion af forbrugstallene i tabellen.
5) Tegn det sæsonkorrigerede forbrugsindeks for forbruget i hele perioden fra 1. kvartal 2000 til 4. kvartal 2005, .
6) Foretag ved hjælp af tendenslinien (regressionslinien) en fremskrivning for de efterfølgende 4 kvartaler i 2006, og beregn på basis af tendenslinien det forventede tal for forbruget af olie i 4. kvartal 2006.

12. Regression

12.1. Indledning

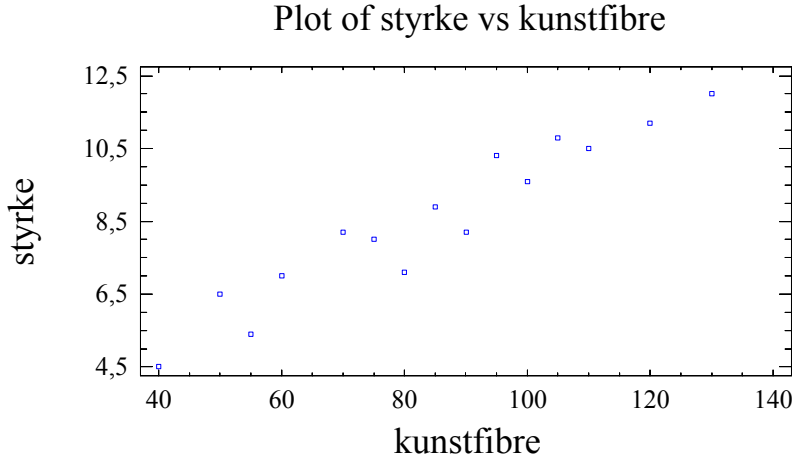
I det forrige kapitel har vi set, hvordan vi kan korrigere for kortvarige sæsonmæssige udsving. Tilsvarende kan man, hvis man har data nok korrigere for de længerevarende konjunktur udsving. Tilbage er der så "trenden", som beskriver den overordnede udvikling over tid.
Det er derfor væsentlig, at man får opstillet et matematisk udtryk for "trenden", idet man så med en passende usikkerhed kan forudsige hvad der vil ske i fremtiden (forudsat naturligvis at udviklingen vil fortsætte som hidtil).
Regressionsanalyse har dog også en række andre anvendelser. Det følgende eksempel viser dette.

Eksempel 12.1 Kvalitet af garn

I et spinderi udtrykkes garnets kvalitet bl.a. ved en norm for den forventede trækstyrke. Kvaliteten anses således for at være i orden, hvis middeltrækstyrken mindst er lig med 10 måleenheder (me).
Ved uldgarn opfylder garnets naturlige trækstyrke ikke det nævnte kvalitetskrav, hvorfor der tilsættes en vis mængde kunstfibre, hvilket forøger trækstyrken. Herved sker der dog det, at andre kvalitetsegenskaber, såsom elasticitet og isoleringsevne, forringes. Man har eksperimenteret med forskellige tilsatte mængder kunstfibre x og registreret garnets trækstyrke y ved disse forskellige mængder. Herved fremkom følgende observationsmateriale:

Mængde x (i gram) af kunstfibre pr. kg uld	40	50	55	60	70	75	80	85	90	95	100	105	110	120	130
Trækstyrke (me): Y	4.5	6.5	5.4	7.0	8.2	8.0	7.1	8.9	8.2	10.3	9.6	10.8	10.5	11.2	12.0

Her er man interesseret i at finde et funktionsudtryk $y = f(x)$ indenfor det måleområde $40 \le x \le 130$ som er af praktisk interesse, mens man ikke er specielt interesseret i at "ekstrapolere", dvs. finde ud af forholdene udenfor området.
Afsættes i eksempel 12.1 de målte punktpar (x_1, y_i) i et koordinatsystem for at få et overblik over forløbet, fås følgende tegning:



Punkterne ligger ikke eksakt på en ret linie, men det synes rimeligt at antage, at afvigelserne fra en ret linie kan forklares ved den tilfældige variation (støjen). ♦

12.2 Regressionslinie og regressionskoefficienter

En matematisk model kunne derfor tænkes at være ligningen for en ret linie $y = a \cdot x + b$, hvor problemet så er, at bestemme konstanterne a og b.
Ligningen kaldes **regressionsligningen** og a og b kaldes **regressionskonstanterne.**

Excel (og mange lommeregnere) kan uden vanskelighed bestemme regressionsligningen.
Ønsker man en nærmere forklaring på beregningerne, sker det i forbindelse med eksempel 12.2, hvor antallet af tal er så (urealistisk) få, at man kan overkomme at foretage beregningerne "i hånden"

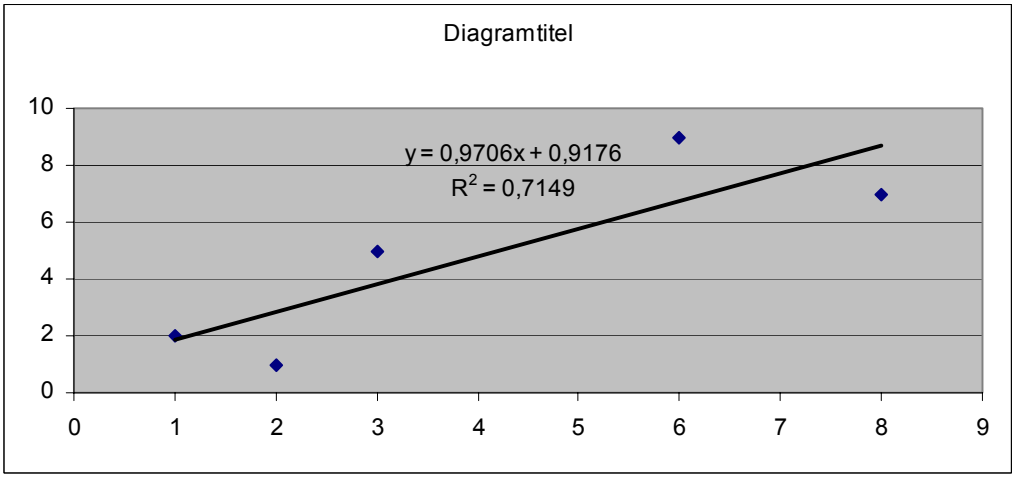
Eksempel 12.2. Bestemmelse af regressionskoefficienter ved mindste kvadraters metode.
I et medicinsk forsøg måles på en forsøgsperson sammenhørende værdier af en bestemt medicin i blodet (i %) og reaktionstiden.
Resultaterne var:

x	1	2	3	6	8
y	2	1	4	9	7

Bestem ved mindste kvadraters metode et estimat for regressionslinien.
Løsning:
Excel:
Data indtastes:

x	y
1	2
2	1
3	5
6	9
8	7

Excel 2003: Marker udskriftsområde ▶ Vælg på værktøjslinien "Guiden Diagram ▶ XY-punkt ▶ Næste ▶ Næste ▶ Næste ▶ Udfør
Placer cursor på et punkt på figuren, højre musetast ▶ Vælg "tilføj tendenslinie" ▶ Vælg indstillinger ▶ vælg"Vis ligning i diagram" og "Vis R-kvadreret i diagram" ▶ ok
Excel 2007: Marker udskriftsområde ▶ Vælg på værktøjslinien "indsæt ▶ Punktdiagram ▶ Vælg "Kun med datomærker ▶ Placer cursor på et punkt på figuren, højre musetast ▶ Vælg "tilføj tendenslinie" ▶ Vælg indstillinger ▶ vælg"Vis ligning i diagram" og "Vis R-kvadreret i diagram" ▶ ok
Der fremkommer følgende figur



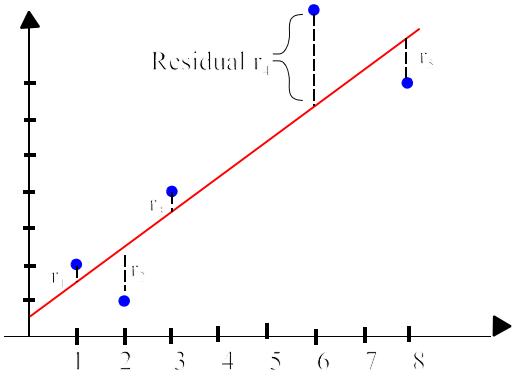
Heraf ses, at regressionsligningen er $y = 0.9706 \cdot x + 0.9176$.

Tallet R^2 kaldes **forklaringsgraden** og er et mål for i hvilken grad regressionsligningen forklarer variationen i tallene. Det gælder, at $0 \le R^2 \le 1$ og jo tættere R^2 er ved 1 jo bedre "passer" ligningen. Ligger punkterne fuldstændigt "diffust" uden nogen sammenhæng siger man at de statistiske variable X og Y er uafhængige. Regressionslinien vil så være næsten vandret ($y \approx \bar{y}$) og $R^2 \approx 0$ Da vi har $R^2 = 0.7149$ vil man løst sagt sige, at linien forklarer 71.5% af variationen.

Forklaring af de indgående størrelser i en regressionsanalyse:

Residual. Ved et punkts residual til en linie forstås den "lodrette" afstand fra punktet til linien (se figur 12.1).

På figur 12.2 er afsat de 5 punkter, og indtegnet en ret linie.



Figur 12.2 *Residualer*

Mindste Kvadraters metode. Regressionslinien bestemmes som den af alle mulige rette linier, for hvilket summen af kvadratet af residualerne til linien er mindst.

I eksempel 12.2 er kvadratsummen SK = $r_1^2 + r_2^2 + r_3^2 + r_4^2 + r_5^2$.

Beregning: I vort tilfælde hvor vi har 5 punkter, indsættes vi disse i ligningen $y = a \cdot x + b$. Dette giver: $2 = a + b \cdot 1$, $1 = a + b \cdot 2$, $4 = a + b \cdot 3$, $9 = a + b \cdot 6$, $7 = a + b \cdot 8$.

Residualerne er så $r_1 = 2-(a+b\cdot 1),\ r_2 = 1-(a+b\cdot 2),\ r_3 = 4-(a+b\cdot 3), r_4 = 9-(a+b\cdot 6), r_5 = 7-(a+b\cdot 8)$
Indsættes disse i SK, får vi en funktion af to variable *a* og *b*.
En sådan funktion kan man ved differentialregning finde en mindste værdi for, og dermed har man fundet de værdier, for hvilken SK er mindst. Man finder her, at a = 0.97 og b = 0.92, dvs. ligningen bliver $y = 0.92 + 0.87 \cdot x$
SK's mindsteværdi kaldes $SK_{residual}$ og kan her beregnes til 12.77
Gennemsnittet af y-værdierne er $\bar{y} = \frac{2+1+5+9+7}{5} = 4.8$
Man beregner nu residualerne til den vandrette linie y = 4.8 og danner kvadratsummen
$SK_{total} = (2-4.8)^2 + (1-4.8)^2 + (5-4.8)^2 + (9-4.8)^2 + (7-4.8)^2 = 44.8$
Forklaringsgraden beregnes nu af $R^2 = 1 - \frac{SK_{residual}}{SK_{total}} = 1 - \frac{12.77}{44.8} = 0.715$.
Hvis punkterne ligger tæt ved regressionslinien vil $SK_{residual}$ være lille i forhold til SK_{total} og dermed vil forklaringsgraden ligge tæt ved 1.
Hvis punkterne ligger fuldstændigt "diffust" fordi der ikke er nogen sammenhæng mellem x og y, vil regressionslinien være (tæt ved) den vandrette linie y = $\bar{y}$. Derved vil de to residualer blive (næsten) lige store, dvs. forklaringsgraden være tæt ved 0. ♦

12.3. Transformation af data inden lineær regressionsanalyse kan foretages.

Ligger punkterne ikke tilnærmelsesvis på en ret linie, er det muligt, at man ved at vælge en passende transformation kan føre problemet over i en lineær model i de transformerede data.
I oversigt 12.3 er angivet en liste med kommentarer over en række muligheder.
Excel har (som andre programmer) indbygget en række transformationer.
Dette illustreres ved følgende eksempel.

Eksempel 12.3 (transformation af udtryk).

Den effekt P (kWatt) som en bil må yde for at overvinde luftmodstanden ved en given hastighed v (km/t) er målt i en vindtunnel. Man fandt følgende sammenhæng mellem v og P.

v	10	30	60	90	120
P	0.01	0.28	2.11	7.35	17.25

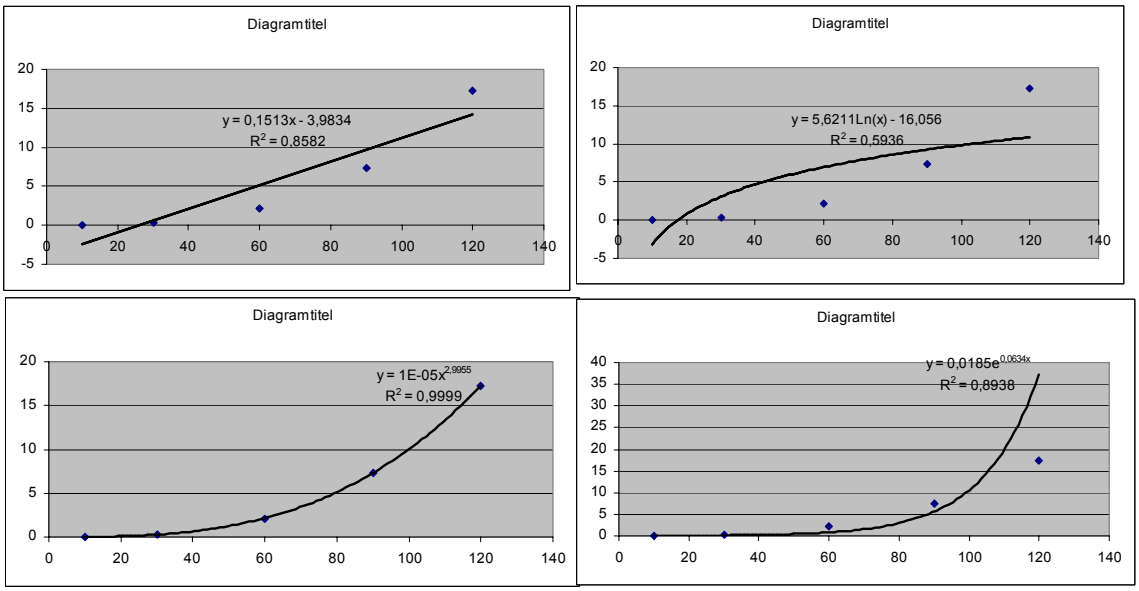
1) Find den model blandt de indbyggede i Excel, som giver den bedste beskrivelse af data.
2) Angiv ligningen for den fundne model
3) Find den effekt der skal ydes ved en hastighed på 100 km.

Løsning:
1) Data indtastes

v	p
10	0,01
30	0,28
60	2,11
90	7,35
120	17,25

Vi indtaster nu på samme måde som i eksempel 12.2, men vælger nu på skift først lineær og hvis den ikke er en god model de øvrige muligheder, dvs., eksponential, logaritmisk og potens model (den sidste hedder i Excel 2007 "strøm")

Resultatet bliver:



Man starter altid med den simpleste model, nemlig en ret linie. Selv om forklaringsgraden er rimelig høj, så ligger punkterne klart ikke på en ret linie. Vi søger derfor efter en anden model.

Det ses, at potensmodellen er den bedste, da den har højest forklaringsgrad, og punkterne ligger også mere "tilfældigt omkring kurven.

2) Ligning $\underline{\underline{P = 10^{-5}\, v^{2.9955}}}$.

3) v = 100 P = =10^(-5)*100^2,9955 = $\underline{\underline{9.79}}$

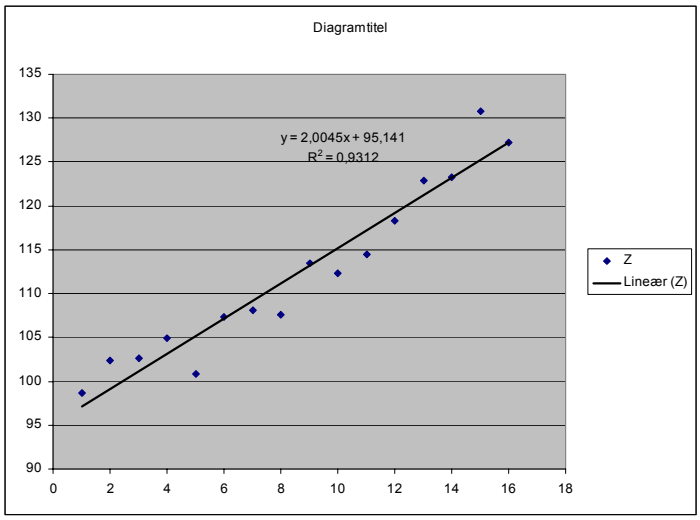
12.4 Trend og fremskrivning ved at benytte regression.

Skal man benytte regression til at komme med en prognose for hvad der vil ske i fremtiden, så skal man naturligvis være klar over, at det forudsætter, at udviklingen fortsætter som hidtil, idet man jo benytter den kendte "historiske" trend til at forudsige fremtiden.
Lad os som eksempel igen vælge de sæsonkorrigerede tal fra afsnit 11.2.5.
Eksemplet findes på adressen i www.larsen-net.dk

Eksempel 12.4 Trend og fremskrivning.

Vi fandt der, at på basis af 4 år (16 kvartaler) følgende tal:

kvartal	Y	Z
02-k1	94	98,69687329
02-k2	103	102,3430267
02-k3	95	102,6497512
02-k4	117	104,8676148
03-k1	96	100,7968068
03-k2	108	107,3111348
03-k3	100	108,0523697
03-k4	120	107,556528
04-k1	108	113,3964076
04-k2	113	112,2792429
04-k3	106	114,5355119
04-k4	132	118,3121808
05-k1	117	122,8461082
05-k2	124	123,2090807
05-k3	121	130,7433673
05-k4	142	127,2752248

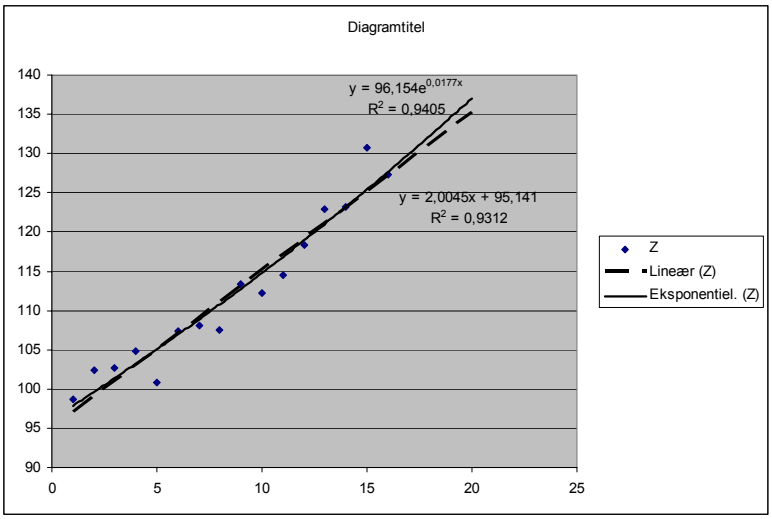


Som det ses, er modellen nogenlunde, men der er dog en tendens til at punkterne i midten ligger under linien og punkterne yderst ligger over.
Vi vælger derfor nu også at se på den eksponentielle vækst. Endvidere fremskriver vi kurverne med 4 kvartaler, for der at se hvor stor forskellen er på de to modeller et år frem.

Excel 2003: Marker udskriftsområde ▶ Vælg på værktøjslinien "Guiden Diagram" ▶ XY-punkt ▶ Næste ▶ Næste ▶ Næste ▶ Udfør
Cursor på Y-akse, højre musetast ▶ Formater akse ▶ skala ▶ minimum = 90 ▶ ok
Placer cursor på et punkt på figuren, højre musetast ▶ Vælg "tilføj tendenslinie" ▶ vælg lineær figur ▶ Vælg indstillinger ▶ vælg "Vis ligning i diagram" , "Vis R-kvadreret i diagram" og sæt prognose 4 enheder frem ▶ ok
Placer cursor på et punkt på figuren, højre musetast ▶ Vælg "tilføj tendenslinie" ▶ vælg eksponentiel figur ▶ Vælg indstillinger ▶ vælg "Vis ligning i diagram" , "Vis R-kvadreret i diagram" og sæt prognose 4 enheder frem ▶ ok

Excel 2007: Marker udskriftsområde ▶ Vælg på værktøjslinien "Indsæt" ▶ Punktdiagram ▶ Vælg "kun med datomærker ▶ Cursor på Y-akse, højre musetast ▶ Formater akse ▶ akseindstilling: Minimum = 90 ▶ ok
Placer cursor på et punkt på figuren, højre musetast ▶ Vælg "tilføj tendenslinie" ▶ vælg lineær ▶ Vælg indstillinger ▶ vælg "Vis ligning i diagram" , "Vis R-kvadreret i diagram" , prognose: Fremad 4 ▶ ok
Placer cursor på et punkt på figuren, højre musetast ▶ Vælg "tilføj tendenslinie" ▶ vælg eksponentiel ▶ Vælg indstillinger ▶ vælg "Vis ligning i diagram" , "Vis R-kvadreret i diagram" og prognose: Fremad 4 ▶ ok

Der fremkommer følgende figur:



Som det ses, er de to kurver indenfor måleområdet stort set ens, men ved fremskrivningen ses, at den eksponentielle kurve stiger stærkest.
Hvilken man skal vælge, afhænger af, om man tror på at y (omsætningen) gennemsnitlig vokser med en konstant absolut størrelse pr kvartel på 2.005 (hældningskoefficienten af den rette linie) eller at den vokser med en konstant gennemsnitlig vækstrate på 0.0177 (samme procentvise vækst).

12.5 Regressionsanalyse

De foregående betragtninger kræver ingen statistiske forudsætninger, idet man jo altid ved mindste kvadraters metode kan beregne regressionskoefficienterne, derefter beregne forklarings-grad, tegne kurver og punkter ind i et koordinatsystem og så herudfra vurdere om modellen er acceptabel.
Lad os antage, at vi har fundet (ved at betragte tegning + forklaringsgrad , at modellen $Y = \alpha + \beta \cdot x$ gælder (*a* og *b* er i det følgende estimater for de eksakte værdier α og β).

I dette afsnit ønsker vi at foretage en nøjere statistisk analyse.

Forudsætninger for regressionsanalyse

Der stilles krav om normalitet og "varianshomogenitet" (se nedenfor punkt 1 og 2), men da de sædvanligvis er opfyldt i praksis, er det unødvendigt at lave ekstra forsøg m.m. for at kunne kontrollere om kravene er opfyldt. Analysen er heldigvis stadig er gyldig, selv om der forekommer mindre afvigelser (testene er "robuste" overfor kravet normalitet og varianshomogenitet)[1].

Test af hældning og opstilling af konfidensintervaller.

Som tidligere nævnt vil en vandret regressionslinie betyde, at punkterne ligger fuldstændigt diffust, dvs. y er uafhængig af x.
Hvis linien er vandret, vil alle x-værdier give samme y-værdi, og tilfældet er ikke interessant.

Vi er derfor interesseret i at kunne teste om hældningskoefficienten β for linien er 0, dvs. nulhypotesen $H_0 : \beta = 0$

Hvis H_0 forkastes, vil vi

2) opstille et konfidensinterval for hældningskoefficienten β_1.
3) for en given værdi for $x = x_0$, dels beregne den tilsvarende "forventede" y_0 værdi (predicted value") dels et konfidensinterval for y_0.

Beregningerne foretages i Excel på tallene fra eksempel 12.1

1

1) De enkelte observationer y_i er indbyrdes uafhængige (eksempelvis hvis der udføres flere målinger for samme mængde medicin skal de være indbyrdes uafhængige, ligesom det også skal gælde målinger baseret på forskellige mængder medicin.
2) Der skal være "varianshomogenitet", dvs. variansen af Y skal være den samme uafhængig af x's værdi.
   Punkt 1 kan opfyldes ved en hensigtsmæssig forsøgsplan. I eksempel 12.2 skal man således være sikker på at den foregående dosis medicin er ude af blodet inden man foretager en ny indsprøjtning, ligesom forsøgene skal være randomiseret. Man kan nok i dette tilfælde betvivle uafhængigheden, hvis man udfører forsøgene på samme person.
   Hvis man er i alvorlig tvivl om kravet om normalitet er rimeligt opfyldt, kan man få et indtryk af, om der er alvorlige afvigelser, ved at tegne et normalfordelingsplot. Man afsætter her residualerne i et koordinatsystem som har den egenskab, at hvis residualerne er normalfordelte vil punkterne (med tilnærmelse) ligge på en ret linie (vises i eksempel 11.7)

Eksempel 12.5 Regressionsanalyse

I et spinderi udtrykkes garnets kvalitet bl.a. ved en norm for den forventede trækstyrke. Kvaliteten anses således for at være i orden, hvis middeltrækstyrken mindst er lig med 10 måleenheder (me).

Man har eksperimenteret med forskellige tilsatte mængder kunstfibre x og registreret garnets trækstyrke y ved disse forskellige mængder. Herved fremkom følgende observationsmateriale:

Mængde x (i gram) af kunstfibre pr kg uld	40	50	55	60	70	75	80	85	90	95	100	105	110	120	130
Trækstyrke (me): Y	4.5	6.5	5.4	7.0	8.2	8.0	7.1	8.9	8.2	10.3	9.6	10.8	10.5	11.2	12.0

1) Vurdér modellen ved at se på forklaringsgraden r^2 , og på en figur med punkterne indtegnet. undersøg endvidere om der kunne tænkes at være "outliers", dvs. punkter der afviger så kraftigt fra linien, at de kunne tænkes at være fejlmålinger.
2) Opskriv regressionsligningen.
3) Test om Y er uafhængig af x
4) Angiv et 95% konfidensinterval for hældningenskoefficienten β
5) Opstil et 95% konfidensinterval for middeltrækstyrken svarende til x - værdien 100.

Løsning:

Af hensyn til de efterfølgende test vælges nu Excels regressionsprogram
Data indtastes på sædvanlig måde i to søjler.

x	y
40	4,5
50	6,5
55	5,4
60	7
osv	osv.
95	10,3
100	9,6
105	10,8
110	10,5
120	11,2
130	12

Excel 2003:Vælg Funktioner ▶ Dataanalyse ▶ Regression
Excel 2007: Data ▶ Dataanalyse ▶ Regression

Den fremkomne tabel udfyldes.
Sædvanligvis er det ikke nødvendigt at udfylde alle felter, men for at kunne forstå de fremkomne udskrifter er dette sket her. Udskrifterne vises dog i en lidt anden rækkefølge.

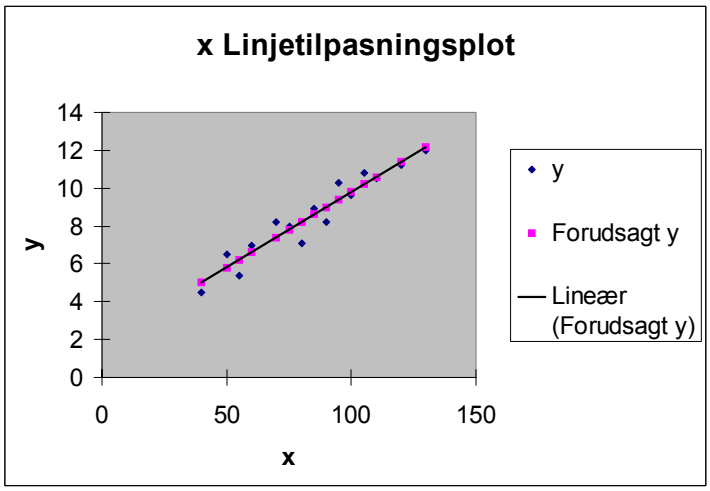
1) Vurdering af model:

RESUMEOUTPUT

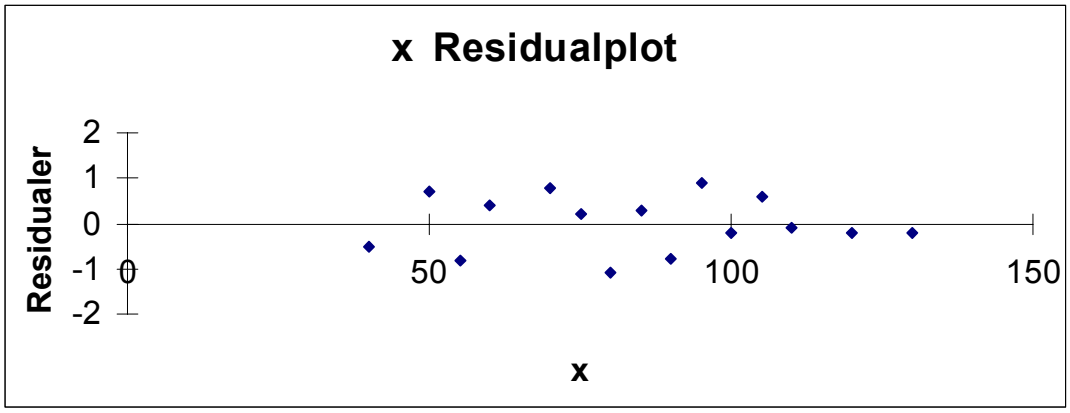
Regressionsstatistik	
Multipel R	0,958802
R-kvadreret	0,919301
Justeret R-kvadreret	0,913093
Standardfejl	0,648068
Observationer	15

Af udskriften ses, at forklaringsgraden "R-kvadreret" er 0.9193. , hvilket er tilfredsstillende, da modellen altså "forklarer" 91,93% af variationen.

Nedenstående plot (som er blevet redigeret i forhold til det oprindelige tegning i Excel) viser, at punkterne fordeler sig tilfældigt omkring linien. Endvidere synes der ikke at være nogen "outliers" (punkter der afviger så kraftigt fra det generelle billede, at man kunne frygte de var fejlmålinger)



Residualplot:
Da linietilpasningsplottet i Excel ofte ikke er særlig overskueligt kan det være bedre at tegne en figur, hvor residualerne er indtegnet (et residualplot)
Af nedenstående plot af residualerne, ses, at de fordeler sig tilfældigt omkring "0 - linien".



Outliers:
Da undersøgelse af outliers er vigtig, har Excel endvidere beregnet såkaldte "Standardresidualer" , hvor man har "normeret" residualerne ,hvorved der tages hensyn til, at konfidensintervallerne er smallest tæt ved "midtpunktet". Ligger en standardresidual numerisk udenfor 2 enheder er der grund til at undersøge om den dertil svarende værdi skulle være en "outliers"

RESIDUALOUTPUT SANDSYNLIGHEDSOUTPUT

Observation	Forudsagt y	Residualer	Standardresidualer	Fraktil	y
1	5,00455	-0,50455	-0,80793	3,333333	4,5
2	5,803524	0,696476	1,115265	10	5,4
3	6,203011	-0,80301	-1,28586	16,66667	6,5
4	6,602497	0,397503	0,63652	23,33333	7
5	7,401471	0,798529	1,278682	30	7,1
6	7,800958	0,199042	0,318725	36,66667	8
7	8,200445	-1,10044	-1,76214	43,33333	8,2
8	8,599932	0,300068	0,480499	50	8,2
9	8,999418	-0,79942	-1,28011	56,66667	8,9
10	9,398905	0,901095	1,44292	63,33333	9,6
11	9,798392	-0,19839	-0,31768	70	10,3
12	10,19788	0,602121	0,964174	76,66667	10,5
13	10,59737	-0,09737	-0,15591	83,33333	10,8
14	11,39634	-0,19634	-0,3144	90	11,2
15	12,19531	-0,19531	-0,31275	96,66667	12

Da standardresidualerne her ligger indenfor 2 enheder på hver side, ses igen at der ikke er nogen outliers.
Modellen synes på tilfredstillende måde at beskrive data.

2) **Regressionsligning:**
Udskrift 1:

	Koeffi-cienter	Stan-dardfejl	t-stat	P-værdi	Nedre 95%	Øvre 95%	Nedre 95,0%	Øvre 95,0%
Skæring	1,808655	0,578421	3,126883	0,008021	0,559052	3,058258	0,559052	3,058258
x	0,079897	0,00657	12,1693	1,77E-08	0,065714	0,094081	0,065714	0,094081

Af udskriften ved "Skæring" aflæses $a = 1.80866$ og ved "x" $b = 0.0798$.
Regressionsligningen bliver derfor $y = 1.80866 + 0.07989 \cdot x$

3) Test af om *Y* er uafhængig af *x* (regressionslinien er vandret)
$H_0 : \beta = 0, \qquad H : \beta \neq 0$

Man beregner en P - værdi.
Dette kan ske på 2 måder.

1) Ud for x i udskrift 1 ses, at P - værdi = $1.7 \cdot 10^{-8}$. Da denne værdi er meget lille (langt under 0.1% forkastes H_0 *Y* er ikke uafhængig af *x*.
eller anderledes udtrykt så er hældningskoefficienten ikke nul (linien er ikke vandret)
2) Man kunne i stedet betragte den "variansanalysetabel", som automatisk bliver udskrevet:

Udskrift 2
ANAVA

	fg	SK	MK	F	Signifikans F
Regression	1	62,19744	62,19744	148,0919	1,77E-08
Residual	13	5,459897	0,419992		
I alt	14	67,65733			

Det ses, ud for "Regression", at $F = 148.09$ og at P-value $= 1.7 \cdot 10^{-8}$, dvs. samme resultat.

Forklaring:

1) Spredningen på hældningen β er $s_\beta = 0.00657$ (ses under standardfejl) .

   Der udføres nu en sædvanlig tosidet t - test på $t = \frac{\beta - 0}{s_\beta} = \frac{0.079897}{0.00657} = 12.1693$ (ses under t-stat), hvor t har N-2 = 13 frihedsgrader.

   Nulhypotesen forkastes, hvis *P*-værdien er under $\frac{\alpha}{2}$. Excel angiver den dobbelte værdi (svarende til begge "haler") så man altid skal sammenligne med signifikansniveauet α

2) Anskuelig forklaring af F - test.
   Hvis modellen gælder så burde punkterne (uanset om H_0 er sand eller ej) ligge eksakt på en ret linie, dvs. kvadratsummen SK af residualerne være 0. s^2_{residual} burde så også være 0. Når det ikke er tilfældet skyldes det "støjen". Et estimat for forsøgsfejlens (støjens) varians σ^2 er derfor s^2_{residual} (= 0.419992, ses under MK i udskrift 2)
   Er H_0 sand, så burde den vandrette regressionslinie være lig med linien $y = \bar{y}$. SK_{total} burde derfor være lig $SK_{residual}$. $SK_{regression} = SK_{total} - SK_{residual}$ burde derfor være 0. $s^2_{\text{regression}}$ burde derfor også være 0. Når det ikke er tilfældet skyldes det, at forsøgsresultaterne har været påvirket af "støjen". Af samme grund som før må derfor også $s^2_{\text{regression}}$ være et estimat for σ^2

   Vi har følgelig, at hvis H_0 er sand, så er $F_{\text{regression}} = \frac{s^2_{\text{regression}}}{s^2_{\text{residual}}} \approx 1$.

   Det kan vises, at hvis nulhypotesen ikke er sand, så vil $F_{\text{regression}} > 1$, og at $F_{\text{regression}}$ er F- fordelt med en tællerfrihedsgrad på 1 og en nævnerfrihedsgrad på N - 2.
   Testen bliver følgelig en ensidet F - test, dvs. H_0 forkastes, hvis P - værdi = $P(F > F_{\text{regression}}) < \alpha$

4) Konfidensinterval for β:

I udskrift 1 ud for *x* under "Nedre 95%" og "Øvre 95%" aflæses $\underline{\underline{[0.0657\,;\,0.0941]}}$

I stedet kunne man benytte formlen for konfidensinterval: $\beta \pm t_{0.975}(N-2) \cdot s_\beta$

Ud for "*x*" under "Standardfejl" står $s_\beta = 0.006565$.

$\beta \pm t_{0.975}(N-2) \cdot s_\beta = 0.0799 \pm 2.16 \cdot 0.006565 = 0.0799 \pm 0.01418 \cdot$ $\underline{\underline{[0.0657\,;\,0.0941]}}$

5) 95% konfidensinterval for middeltrækstyrken svarende til *x* - værdien 100.

Excel har desværre ikke beregnet dette direkte, så vi må benytte formlen til beregningen.

Konfidensinterval for y_{100} : $y_{100} \pm t_{0.975}(15-2) \cdot \sqrt{V(y_{100})}$ hvor $V(y_{00}) = s^2_{residual} \cdot \left(\frac{1}{N} + \frac{(\beta(x - \bar{x}))^2}{SAK_{\text{regression}}} \right)$

Nedenstående excel - eksempel findes på adressen i www.larsen-net.dk

	A	B	C	D	E	F	G	H
1	x	y	RESUMEOUTPUT					
2	40	4,5						
3	50	6,5	*Regressionsstatistik*					
4	55	5,4	Multipel R	0,958802				
5	60	7	R-kvadreret	0,919301				
6	70	8,2	Justeret R-kvadreret	0,913093				
7	75	8	Standardfejl	0,648068				
8	80	7,1	Observationer	15				
9	85	8,9						
10	90	8,2	ANAVA					
11	95	10,3		*fg*	*SK*	*MK*	*F*	*Signifikans F*
12	100	9,6	Regression	1	62,19744	62,19744	148,0919	1,77E-08
13	105	10,8	Residual	13	5,459897	0,419992		
14	110	10,5	I alt	14	67,65733			
15	120	11,2						
16	130	12		*Koefficient*	*Standardfejl*	*t-stat*	*P-værdi*	*Nedre 95%*
17			Skæring	1,808655	0,578421	3,126883	0,008021	0,559052
18			x	0,079897	0,006565	12,1693	1,77E-08	0,065713
19								
20	**Konfidensinterval for x =**			100				
21	n =	15						
22	y(100) =	D17+D18*D20 =		9,798392				
23								
24	r =	KVROD(F13*(1/b21+(D18*(D20-MIDDEL(A2:A16))^2)/E12))		0,424332				
25	nedre grænse			9,37406				
26	øvre grænse =			10,22272				

De enkelte størrelser i exceludskriften beregnes på følgende måde:

a) Af regressionsligningen $y = 1.80866 + 0.07989 \cdot x$ fås ved indsættelse af $x = 100$ at et estimat for middelværdien $y_{100} = 0.0799 \cdot 100 + 1.8087 = 9.80$

b) $MK_{residual} = 0.419992$ (ses af ANAVA-Tabel),

c) N = observationer =15 (antal punkter),

d) , $SK_{\text{regression}} = 62.19744$ (ses ud for Regression i ANAVA-tabel) og $\beta = 0.079897$.

e) $\bar{x}$ er gennemsnittet af x - værdierne. $\bar{x}$ =MIDDEL(A2:A16) =84,33333

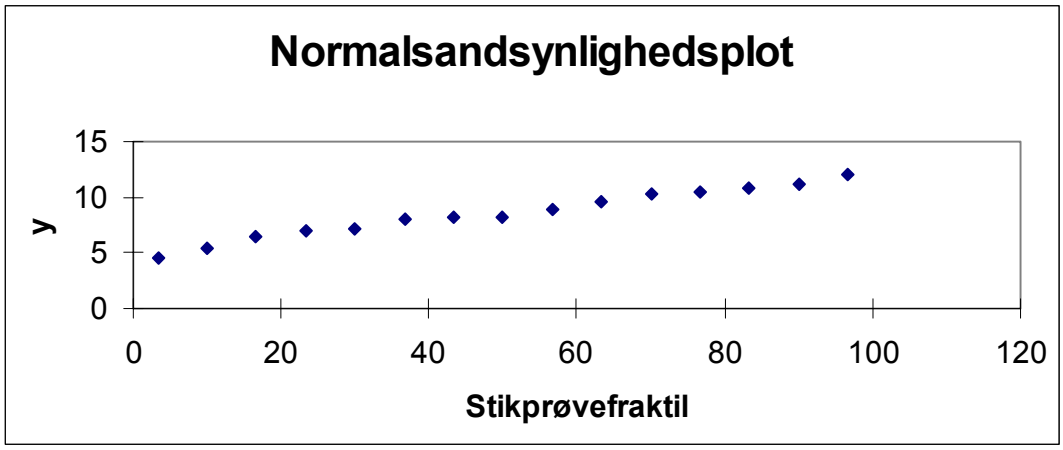
$$\tilde{V}(y_{100}) = MK_{\text{residual}} \cdot \left(\frac{1}{\text{observationer}} + \frac{(\beta_1 (x - \bar{x}))^2}{SK_{\text{regression}}} \right) = 0.4200 \left(\frac{1}{15} + \frac{(0.079897(100 - 84.33)9^2}{62.19744} \right) = 0.0386 .$$

$$\hat{\mu} \pm t_{0.975}(15-2) \cdot \sqrt{\hat{V}(\hat{\mu})} = 9.80 \pm 2.16 \cdot \sqrt{0.0386} = 9.80 \pm 0.4243. \quad \underline{\underline{[9.38;10.22]}}$$

◇

Ønskes en kontrol af forudsætningen om at residualerne skal være normalfordelt kunne man have valgt "plot af normal sandsynlighed".

Man får så følgende tegning:



Da residualpunkterne på ovenstående normalfordelingsplot tilnærmelsesvis ligger på en ret linie antages forudsætningen om normalitet at være opfyldt. ◇

Opgaver

Opgave 12.1
I opgave 11.1 blev beregnet et sæsonkorrigeret mængdeindeks for en virksomheds omsætning i perioden 1. kvartal 2002 til 1 kvartal 2006.
1) Foretag ved hjælp af regression (tendenslinier) en fremskrivning for de efterfølgende 3 kvartaler i 2006. De forskellige muligheder for tendenslinier skal sammenlignes.
2) Beregn på basis af den model der foretrækkes i spørgsmål 1 en beregning af det forventede tal for omsætningen i 4. kvartal i 2006.

Opgave 12.2
I opgave 11.2 blev beregnet et sæsonkorrigeret indekstal for perioden 1 kvartal 2002 til 1 kvartal 2006.
1) Foretag ved hjælp af regression (tendenslinier) en fremskrivning for de efterfølgende 3 kvartaler i 2006. De forskellige muligheder for tendenslinier skal sammenlignes.
2) Beregn på basis af den model der foretrækkes i spørgsmål 1 en beregning af det forventede tal for kvartalsindeks i 4 kvartal 2006.

Opgave 12.3
I et forsøg undersøgtes et ventilationsanlægs effektivitet. Målingerne foretoges ved at fylde et lokale med gas og vente til koncentrationen var stabil. Herefter startedes ventilationsanlægget og gaskoncentrationen C måltes til forskellige tidspunkter t.
Følgende resultater fandtes:

t (min. efter anlæggets start)	2.67	4.59	6.75	7.67	11.34	14.34	16.25	18.25	23.09
C [ppm]	34	28	26	22	16	14	12	10	8

Følgende 2 modeller for funktionssammenhængen overvejes:
Model 1 (lineært henfald): $C = a + b \cdot t$
Mode1 2 (eksponentielt henfald): $C = a \cdot e^{b \cdot t}$
1) Indtegn punkterne i et koordinatsystem og vælg på basis af tegning og forklaringsgrad den af de to modeller du vurderer giver den bedste beskrivelse.
2) Opskriv regressionsligningen for den i spørgsmål 1) fundne model.
3) Bestem for den i spørgsmål 1) fundne model værdien af C for t = 10.0.

Opgave 12.4

Man mener der er en sammenhæng mellem en bilists alder og antallet af alvorlige færdselsulykker, der skyldes for stor hastighed. Man har fra USA, hvor aldersgrænsen for erhvervelse af kørekort er 16 år, følgende data indsamlet gennem en periode:

Alder x	16	17	18	19	20	22	24	27	32	42	52	57	62	72
Antal fart-relaterede ulykker y	37	32	33	34	33	31	28	26	23	16	13	10	9	7

Det fremgår klart, at antallet af ulykker falder med alderen.

1) Giv en vurdering af, om modellen : $y = \alpha_0 + \beta_1 x$ (antal ulykker aftager lineært med alderen) på rimelig måde kan beskrive denne sammenhæng
2) En trafikekspert mener, at modellen $y = \alpha_1 \cdot e^{\beta_2 x}$ (antal ulykker aftager eksponentielt med alderen) giver en bedre beskrivelse af modellen. Har vedkommende ret?
3) Bestem Regressionsligningen for den model du finder bedst beskriver data.
4) Angiv det forventede antal fart-relaterede ulykker som 50 - årige i middel vil forårsage i den givne periode.

Opgave 12.5

Følgende sammenhørende data er 25 målinger mellem den jævnstrøm (y) en vindmølle udvikler og vindhastigheden (x). Data findes på adressen www.larsen-net.dk

x	5.00	6.00	3.40	2.7	10	9.7	9.55	3.05	8.15	6.2	2.90	6.35	4.60
y	1.582	1.822	1.057	0.500	2.236	2.386	2.294	0.588	2.166	1.866	0.653	1.930	1.562
x	5.80	7.40	3.60	7.85	8.80	7.00	5.45	9.10	10.20	4.10	3.95	2.45	
y	1.737	2.088	1.137	2.179	2.112	1.800	1.501	2.303	2.310	1.194	1.144	0.123	

1) Angiv (med begrundelse) den af de i Excel angivne modeller "Lineær, potens,eksponentiel, polynomisk", som du finder bedst beskriver y's variation som funktion af x.

For den fundne model skal man

2) angive ligningen for den fundne sammenhæng mellem x og y
3) Før kurven frem til en vindhastighed på x = 12 og aflæs (approksimativt) (på tegningen) værdien af y for x = 12
5) Beregn ud fra ligningen værdien af y svarende til x = 12

Appendix A. Oversigt over Excel-kommandoer

A: Generelle forhold

a1: Forudsætninger.

Da ikke alle de anvendte statistiske funktioner er indbygget fra starten, skal man først vælge et tilføjelsesprogram:
I Excel 2003: Vælg "Funktioner", "Tilføjelsesprogrammer", marker "Problemløser"
I Excel 2007: Vælg "Excel-Office-knappen", "Excel indstillinger (findes forneden)", Tilføjelsesprogrammer", "Udfør", "marker Problemløser", "Installer".

a2. Inddata.

Placering: Vi vil i det følgende for kortheds skyld antage, at den første stikprøves værdier står i cellerne A1, A2, A3 . . . A10. Kræves der flere variable vil den næste stå i cellerne B1, B2, B3 . . . B8, osv.

Man angiver "udskriftsområdet" eller "inputområdet" f.eks en søjle placere i cellerne A1:A10 ved
a) at markere området A1 til A10
b) at skrive eksempelvis A1:A10
c) at give det et navn: Vælg "Indsæt" ▶ i Excel 2003: Navn i Excel 2007:Formler ▶ Definer ▶ i menu skriv søjlens navn og (nederst)A1:A10

a3. Skrive , beregne og kopiere formler.

Vælg den celle hvor resultatet skal stå. Lad det være B1: ▶ På værktøjslinien foroven skriv = ▶ formel skrives ▶ ENTER Resultatet står nu i celle B1
Hvis selve formlen skal stå i en anden celle. Lad det være A1: Cursor placeres i B1 ▶ I formelfelt markeres formlen uden lighedstegn og man kopierer den (CTRL C)" ▶ ENTER (så formlen igen er beregnet i B1 ▶ Cursor over i A1 og paste (CTRL V)

a4. Udskrive gitterlinier og række og kolonneoverskrifter

Excel 2003: Vælg Filer ▶ Sideopsætning ▶ Ark ▶ Marker gitterlinier ▶ marker række- og kolonneoverskrifter.
Excel 2007: Vælg Sidelayout ▶ Under"Gitterlinier" marker "Udskriv" ▶ Under "Overskrifter" marker "Udskriv"

a5. Indsætte græske bogstaver

Excel 2003: Vælg Indsæt ▶ symbol ▶ Vælg Skrifttype = Times roman og Undersæt = Græsk standard
Excel 2007: Vælg Indsæt ▶ symbol ▶ Vælg Skrifttype = Times roman og Undersæt = Græsk og koptisk

B: Indsætte og tegne diagrammer

b1. Indsætte diagrammer

Lagkage eller søjle (se evt eksempel 1.3 side 3)
Excel 2003: Marker udskriftsområde ▶ Vælg på værktøjslinien "Guiden diagram" ▶ Cirkel (eller søjle) ▶ Marker ønsket figur ▶ Næste ▶ Navn på kategori ▶ Udfør
Excel 2007: Marker udskriftområde ▶ Vælg på værktøjslinien "Indsæt" ▶ Cirkel (eller søjle ▶ Marker ønsket figur
Kurve: (se evt eksempel 1.4 side 4)
Som ovenfor, men Excel 2003: Vælg kurve
Excel 2007: Vælg Streg

b2. Tegne histogram (se eksempelvis eksempel 1.5 side 6)

Data indtastes i eksempelvis søjle A1 til A10
Excel 2003: Vælg "Funktioner", Dataanalyse, Histogram
Excel 2007: Vælg "Data", Dataanalyse, Histogram
I den fremkomne tabel udfyldes "inputområdet" med A1:A10 og man vælger "diagramoutput"..
1) Trykkes på OK fås en tabel med hyppigheder, og en figur, hvor intervalgrænserne er fastlagt af Excel.

2) Ønsker man selv at bestemme grænserne, skal man også udfylde intervalområdet. Dette gøres ved at skrive de øvre grænser i en søjle (f.eks i B1 -18.7, i B2 -12.9 osv) og så skrive B1:B11 i inputområdet

Nedenstående figurer er blevet gjort lidt "pænere" ved
cursor på en søjle ▶ tryk højre musetast ▶ formater dataserie ▶ indstilling ▶ mellemrumsbredde = 0 ▶ ok

b3. Tegne sumpolygon (se eksempelvis eksempel 1.5 side 6)

Data indtastes i eksempelvis søjle A1 til A10
Excel 2003: Vælg "Funktioner", Dataanalyse, Histogram
Excel 2007: Vælg "Data", Dataanalyse, Histogram
I den fremkomne tabel udfyldes "inputområdet" med A1:A10 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.

C: Beregne statistiske størrelser og funktioner

c1. Beregning af "Karakteristiske tal" (se evt. side 9)

Data indtastes i eksempelvis søjle A1 til A10
Excel 2003: Funktioner ▶ Dataanalyse ▶ Beskrivende statistik ▶ udfyld inputområde ▶ Resumestatistik
Excel 2007: Data ▶ Dataanalyse ▶ Beskrivende statistik ▶ udfyld inputområde ▶ Resumestatistik

c2. Valg af statistiske størrelser (funktioner)

1) Vælg den celle hvor resultatet skal stå (eksempelvis A1).
2) På værktøjslinien foroven:
   2a) Tryk på f_x
   2b) På den fremkommne menu vælges den ønskede funktion eksempelvis "NORMALFORDELING"
   2c) Der fremkommer en menu med anvisning på, hvordan den skal udfyldes.

c3. Gennemsnit, spredning, median, kvartil

Navnene anføres nedenunder, men den fremkomne menu gør det let at indsætte de rette parametre. (de anføres dog her)

Gennemsnit $\bar{x}$ = MIDDEL(A1:A10)
Spredning s = STDAFV (A1:A10)
Median m = MEDIAN(A1:A10) (= KVARTIL(A1:A10;2))
1. Kvartil = KVARTIL(A1:A10;1)

c4. Fakultet, kombination, Permutation (se evt. side 50)

Fakultet n! = FAKULTET(n) Eksempel: 5! =FAKULTET(5) = 120
Kombination K(n,p) = KOMBIN(n;p) Eksempel: K(5,3)==KOMBIN(5;3) = 10
Permutation P(n,p = PERMUT(n;p) Eksempel: P(5,3) = PERMUT(5;3) = 60

c5. Normalfordeling. (se evt. side 24)

Lad X være normalfordelt med middelværdi μ og spredning σ

1) $P(X \le x)$ = NORMFORDELING(x ; μ ; σ ;1)
2) $P(X \ge x)$ = 1 - NORMFORDELING(x ; μ ; σ ;1)
3) $P(a \le X \le b) = P(X \le b) - P(X \le a)$ =
   NORMFORDELING(b ; μ ; σ ;1) -NORMFORDELING(a ; μ ; σ ;1)

Fraktil x_p

$P(X \le x_p) = p \Leftrightarrow x_p$ = NORMINV(p; μ ; σ)

Eksempel: $u_{0.975}$ = NORMINV(0,975;0;1) = 1,959961

c6. *t* - fordeling. (se evt. side 33 og 91)

Lad T være t - fordelt med f frihedsgrader..

1) $P(T \ge |t|)$ = TFORDELING(abs(t); f ;1)
   (bemærk: TFORDELING(abs(t); f ;1) udregner "øvre hale" af fordelingen)
2) $P(T \le -|t|) + P(T \ge |t|)$ = TFORDELING(abs(t); f ;2) (udregner "halen" til begge sider)

Fraktil

$t_\alpha(f)$ = TINV(2(1 - α) ; f) , $\alpha > 0.5$

$t_\alpha(f)$ = - TINV(2 α ; f) , $\alpha < 0.5$

Bemærk: TINV(α ; f) udregner "øvre hale", svarende til 1 - $\frac{\alpha}{2}$

Bemærk: Man må må udnytte symmetrien i t-fordelingen, for værdier mindre end 0 (svarende til $\alpha < 0.5$)

Eksempel:

Lad T være t - fordelt med 12 frihedsgrader

1) $P(X \le -1) = P(X \ge 1)$ = TFORDELING(abs(-1);12;1) = 0,168525
2) $t_{0.975}(12)$ = TINV(0,05;12) = 2,178813
   $t_{0.025}(12)$ = - TINV(0,05;12) = - 2,178813

c7. χ^2 - fordeling. (se side 33)

Lad X være χ^2 - fordelt med f frihedsgrader

$P(X \ge x)$ = CHIFORDELING(x;f)

(bemærk: CHIFORDELING(x;f) udregner "øvre hale" af fordelingen)

Fraktil

$\chi^2_\alpha(f)$ =CHIINV(1- α ;f)

(bemærk: CHIINV(α ;f) udregner "øvre hale")

Eksempel:

Lad X være χ^2 - fordelt med 8 frihedsgrader

1) $P(X \le 5)$ = 1- CHIFORDELING(5;8) = 0,242424
2) $\chi^2_{0.975}(8)$ =CHIINV(0,025;8) = 17,53454
   $\chi^2_{0.025}(8)$ =CHIINV(0,975;8) = 2,179725

c8. *F* - fordeling.

Lad X være F - fordelt med tællerfrihedsgrader på f_T og nævnerfrihedsgrader på f_N

$P(X \ge x)$ = FFORDELING(x ; f_T ; f_N)

Fraktil

$F_\alpha(f_T, f_N)$ ==FINV(1 - α ; f_T ; f_N)

c9. Hypergeometrisk fordeling (se evt. side 53)

Lad X være hypergeometrisk fordelt med parametrene N, M og n

$P(X = x)$ = HYPGEOFORDELING(x ; n ; M ; N)

Eksempel: Lad N = 600, M = 10 og n = 25

$P(X \le 1)$ = HYPGEOFORDELING(1;25;10;600)+HYPGEOFORDELING(0;25;10;600) = 0,938876

c10. Binomialfordeling (se evt. side 58)

Lad X være binomialfordelt med parametrene n og p

$P(X = x)$ = BINOMIALFORDELING(x ; n; p; 0)

$P(X \le x)$ = BINOMIALFORDELING(x ; n; p; 1)

Eksempel (jævnfør eksempel 72)

Lad X være binomialfordelt med n = 6 og p = 0.15

$P(X = 3)$ = BINOMIALFORDELING(3;6;0,15;0) = 0,041453

$P(X \ge 3)$ = 1- $P(X \le 2)$ =1 - BINOMIALFORDELING(2;6;0,15;1) = 0,047339

c11. Poissonfordeling (se evt. side 69)

Lad X være Poissonfordelt med middelværdien μ

$P(X = x)$ = POISSON(x; μ ; 0)

$P(X \le x)$ = POISSON(x; μ ; 1)

Eksempel

Lad X være Poissonfordelt med middelværdien 10

$P(X = 4)$ = POISSON(4; 10;0) = 0.018917

$P(X \ge 4)$ = 1 - POISSON(4;10;1) = 0,970747

c12. Eksponentialfordeling (se evt. side 71)

Lad T være eksponentialfordelt med middelværdien μ.

$P(T \le t)$ = EKSPFORDELING(t,1/ μ ,1)

Eksempel:

Lad T være eksponentialfordelt med middelværdi μ=2

$P(T \le 2)$ = EKSPFORDELING(3;1/2;1) = 0,77687

D: Køteori

Begrænset antal ventende kunder: se exceludskrift i eksempel 9.2 side 79

Ubegrænset antal ventende kunder se exceludskrift i eksempel 9.3 side82

E: Konfidensintervaller

e1. Konfidensinterval middelværdi for 1 normalfordelt variabel. σ kendt eksakt

Radius r i et 95% konfidensinterval for μ : $\bar{x} \pm r = \bar{x} \pm u_{0.975} \frac{\sigma}{\sqrt{n}}$ (se evt. side 32)

r = KONFIDENSINTERVAL(0,05; σ , n).

Eksempel. Lad stikprøven have n =6 værdier, lad spredning σ = 0.25 og gennemsnit $\bar{x}$ =8

r =KONFIDENSINTERVAL(0,05;0,25;6). Resultat 0,200038

95% konfidensinterval: 8,0 $\pm$ 0.200

e2. Konfidensinterval for middelværdi for 1 normalfordelt variabel . σ ikke kendt eksakt

Radius r i et 95% konfidensinterval for $\mu : \bar{x} \pm r = \bar{x} \pm t_{0,975}(f)\frac{s}{\sqrt{n}}$:

1) Data givet: Lad data være tallene fra eksempel 4.5 side 34. De står i cellerne A1 til A7 .
Vælg "Funktioner", Dataanalyse, Beskrivende statistik. I den fremkomne tabel udfyldes "inputområdet" med A1:A6, Outputområdet til B1, og konfidensniveau for middelværdi sættes til 95%
Vælg den celle, hvor konfidensintervallets radius skal stå

Resultat:

	Kolonne1
Konfidensniveau(95,0%)	12,7589

$\bar{x}$ = Middel (A1:A7)= 49.08 r = 12.7589 95% konfidensinterval: 49.08 ± 12.7589

2) Hvis data ikke er kendt, men kun n, $\bar{x}$ og s , må formel benyttes. Se eksempel 4.6 side 35

e3. 95%- konfidensinterval for spredning af stikprøve idet μ er ukendt

se eksempel 4.8 side 37

e4. Konfidensinterval for sandsynlighed *p* for 1 binomialfordelt variabel.

se Excel-udskrift i eksempel 7.6 side 61

F: Hypotesetest

f1. 1 normalfordelt variabel

se Excel-program i eksempel 10.4 side 92

f2. 2 normalfordelte variable

1) Ikke parvise observationer:
data givet: se Excel-program i eksempel 10.5 side 95
data ikke givet: se Excel-program i eksempel 10.6 side 96
2) Parvise observationer:
se Excel-program i eksempel 10.7 side 98

f3: 1 binomialfordelt variabel

se oversigt 10.1 + eksempel 10.2 side 88

f4: 2 binomialfordelt variabel

se Excel-program i eksempel 10.3 side 90

G. Antalstabeller

g1. En - vejs tabel

Skriv de observerede værdier i en søjle f.eks A1- A4 og de forventede værdier i en anden søjle f.eks B1 - B4
Eksempel: (hentet fra eksempel 10.8 side 102)

Inddata:

39	35
336	325
99	90
26	50

Skriv =CHITEST(A1:A4;B1:B4) P - værdi = 0,004127

g2. To - vejs tabel

Skriv de observerede værdier i et område f.eks A1 - D4 og de forventede værdier i eksempelvis A6 - D9
Eksempel: (hentet fra eksempel 10.11 side 103)

18	46	13	0
22	60	42	5
7	123	42	16
2	28	68	8

7,546	39,578	25,410	4,466
12,642	66,306	42,570	7,482
18,424	96,632	62,040	10,904
10,388	54,484	34,980	6,148

skriv = CHITEST(D1:G4;D6:G9) Resultat: P - værdi = 2,44228E-19

H: Tidsrækker.

Multiplikativ og additiv metode se eksempel 11.1 sde 118 - 123

I. Regression

i 1. Bestemmelse af regresionskoefficient og regressionslinie (se evt. eksempel 12.2 side 124)

Data indtastes:

Excel 2003: Marker udskriftsområde ▶ Vælg på værktøjslinien "Guiden Diagram ▶ XY-punkt ▶ Næste ▶ Næste ▶ Næste ▶ Udfør
Placer cursor på et punkt på figuren, højre musetast ▶ Vælg "tilføj tendenslinie" ▶ Vælg indstillinger ▶ vælg"Vis ligning i diagram" og "Vis R-kvadreret i diagram" ▶ ok

Excel 2007: Marker udskriftsområde ▶ Vælg på værktøjslinien "indsæt ▶ Punktdiagram ▶ Vælg "Kun med datomærker ▶ Placer cursor på et punkt på figuren, højre musetast ▶ Vælg "tilføj tendenslinie" ▶ Vælg indstillinger ▶ vælg"Vis ligning i diagram" og "Vis R-kvadreret i diagram" ▶ ok

i 2. Transformation

Samme ordrer som i i 1), men vælger nu på skift først lineær og hvis den ikke er en god model de øvrige muligheder, dvs., eksponential logaritmisk og potens model (den sidste hedder i Excel 2007 "strøm")

i3. Fremskrive regressionslinien (se evt. eksempel 12.4 side 128)

Excel 2003: Marker udskriftsområde ▶ Vælg på værktøjslinien "Guiden Diagram" ▶ XY-punkt ▶ Næste ▶ Næste ▶ Næste ▶ Udfør
Cursor på Y-akse, højre musetast ▶ Formater akse ▶ skala ▶ minimum = 90 ▶ ok
Placer cursor på et punkt på figuren, højre musetast ▶ Vælg "tilføj tendenslinie" ▶ vælg lineær figur ▶ Vælg indstillinger ▶ vælg"Vis ligning i diagram" , "Vis R-kvadreret i diagram" og sæt prognose 4 enheder frem ▶ ok

Excel 2007: Marker udskriftsområde ▶ Vælg på værktøjslinien "Indsæt" ▶ Punktdiagram ▶ Vælg "kun med datomærker ▶ Cursor på Y-akse, højre musetast ▶ Formater akse ▶ akseindstilling: Minimum = 90 ▶ ok
Placer cursor på et punkt på figuren, højre musetast ▶ Vælg "tilføj tendenslinie" ▶ vælg lineær ▶ Vælg indstillinger ▶ vælg"Vis ligning i diagram" , "Vis R-kvadreret i diagram" , prognose: Fremad 4 ▶ ok

i4) Regressionsanalyse

Excel 2003:Vælg Funktioner ▶ Dataanalyse ▶ Regression
Excel 2007: Data ▶ Dataanalyse ▶ Regression
Den fremkomne menu udfyldes

1) I outputområdet skrives øverste venstre celle i det ønskede outputområde.
   Der skal bruges mindst 7 kolonner til regressionsanalysetabellen.
2) Det kan være hensigtsmæssigt at afkrydse yderligere nogle rubrikker, men det er ikke obligatorisk.

Eksempel med tilhørende udskrift kan ses i eksempel 12.5 side 131-135.
Specielt skal bemærkes Excel-program til beregning af konfidensinterval side 135.

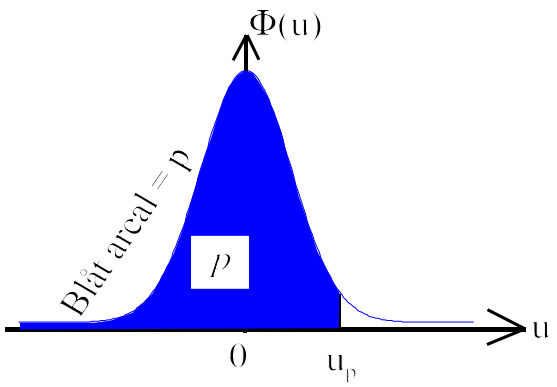
Tabel 1 Fraktiler u_p i U-fordelingen $n(0,1)$. $P(U \le u_p) = p$.

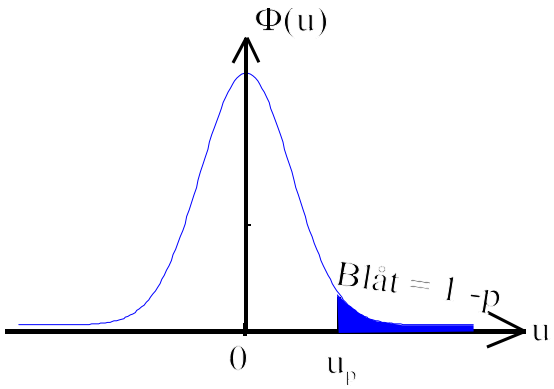
Bemærk: $u_p = - u_{1-p}$

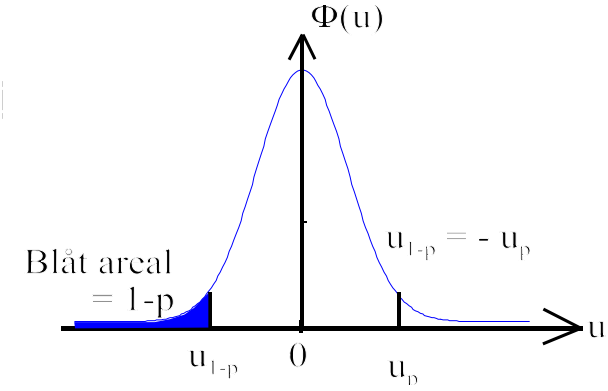
p	0.0005	0.001	0.005	0.01	0.025	0.05	0.10
u_p	-3.291	-3.090	-2.576	-2.326	-1.960	-1.645	-1.282

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 -
1.2 -
1.3 -
1.4 -
1.5 (1) - (2) ca 24%
1.6 (1) - (2) ca 0.052
1.7 (1) - (2) ca 13 %
1.8 (1) - (2) relativ kvartilafsand ca 0.12
1.9 (1) - (2) - (3) - (4) relativ kvartilafstand 0.12 og 0.18 , Ja

Kapitel 3

3.1 (1) 0.7734 0.0548 0.1718 (2) 0.7480
3.2 (1) 69.15% (2) 10.88% (3) 114.4 (4) 117.3 6.535
3.3 78.87%
3.4 (1) 86.64% (2) 0.008 (3) 2.48 2.52
3.5 (1) 5.91% (2) 27.64% (3) [783.51; 816.49]
3.6 (1) 13 (2) 92.8%
3.7 (1) 2.28% (2) 0.078%
3.8 (1) 9.5 1.265 (2) 12.44 (3) 2.41%

Kapitel 4

4.1 (1) 2259.92 35.569 (2) [2237 ; 2283] (3) 25
4.2 (1) 74.0362 0.00124 (2) [74.035; 74.037] (3) 43
4.3 (1) 8.268 0.241 (2) [8.02 ; 8.52] (3) 27
4.4 [4.23 ; 4.29]
4.5 0.965 ; 1.111]
4.6 (1) 367 81 8 (2) - (3) nej (4) 1100 24 (5) 145 103
4.7 (1) - (2) - (3) - (4) - (5) [938 ; 1056] (6) -
4.8 [25.21; 60.36]
4.9 [0.00083 ; 0.00231]
4.10
4.11

Kapitel 5

5.1 0.1 0.5 0.8 0.2 0.7
5.2 (1) 0.9134 (2) 0.9678
5.3 (1) 8.75% (2) 38.75% (3) 41.25% (4)11.25%
5.4 (1) 6.4% (2) 78.4% (3) 7.2%
5.5 (a) 30.24% (b) 0.24% (c) 99.76% (d) 4.04% (e) 44.04% (f) 21.44%
5.6 (1) 27.1% 36.0% 9.756% (2) 53.34% (3) 49.20%
5.7 (1) - (2) 5 3.75 4.082 4.437 (3) 41.67% (4) 12%

Kapitel 6

6.1 (a) 6 (b) 24
6.2 (1) 100 (2) 2400
6.3 $1.283 \cdot 10^{12}$
6.4 (a) - (b) 736
6.5 30
6.6 30.65%
6.7 (1) 2.38% (2) 73.8%
6.8 17.1%
6.9 (A) 0.018% (B) 1.29% (C) 38.24%
6.10 44.57%
6.11 (1) 0.435% (2) 49.57% (3) 41.30%
6.12 (1) 91.67% (2) 25.00% (3) 9.167%
6.13 (1) 17.68% (2) 59.28%
6.14 3^{40}
6.15 31
6.16 30.24%
6.17 $9 \cdot 10^{7}$
6.18 24.55%

Kapitel 7

7.1 (a) 27.87% (b) 68.46%
7.2 (a) 34.10% (b) 40
7.3 37.11%
7.4 0.0275%
7.5 (a) 9.05% (b) 25
7.6 19.77%
7.7 (1) - (2) - (3) - (4) 45 (5) 70.68% (6) 40.12%
7.8 94.5%
7.9 77.86%
7.10 5.83%
7.11 12.85%
7.12 (1) 7.94% (2) 11.8%
7.13 (1) 0.108 (2) [0.089 ; 0.127] (3) 21.04
7.14 (1) [0.30 ; 0.38] (b) 784
7.15 (1) [0.03 ; 0.07] (2) ca 1825
7.16 (1) [0.04;0.10] (2) ca 625
7.17 (1) 0.683 (2) [0.600 ; 0.767] (3) 322
7.18 [0.799 ; 0.847]
7.19 [0.0048 ; 0.0202]

Kapitel 8

8.1 (1) 37.12% (2) 1.83%
8.2 (1) 91.61% (2) 22.80%
8.3 50.37%
8.4 75.3%
8.5 6.56%
8.6 93.19%
8.7 44.6%
8.8 (1) 29.2% (2) 34.82% (3) 89.65% (4) 7.83
8.9 (1) 15 (2) 81.9%
8.10 (1a) 11 (1b) [4.5 ; 17.5] (2a) 1.1 (2b) [0.45 ; 1.75]
8.11 (1) 30.1% (2) 87.9% (3) 4
8.12 (1) 39.35% (2) 18
8.13 (a) 73.64% (b) 51.34% (c) 74.23% (d) 25.77% (e) 3.67% (f) 21
8.14 (1) 2.90 (2) 14.6% (3) 17
8.15 (1) 77.88% (2) 10.45% (3) 77.88% (4) 9.48%
8.16 (1) 79.8% (2) 99.33% (3) 22.8 (4) 14

Kapitel 9

9.1 ja
9.2 (a) 18.8% (b) 6.5 (c) 109 minutter (d) 78.7% (e) 0.187%
9.3 (a) 7.6% (b) 1.85 (c) 0.23 døgn
9.4 Ja
9.5 (a) 20% (b) 33.3%
9.6 (a) 0.5 (b) 50% (c) ja
9.7 (a) 33,3% (b) 2
9.8 Nej
9.9 (a) 1.52 3 (b) 0.25 minutter
9.10 1.93
9.11 (a) 1.05 (b) 9.16% (c) 62.5% (d) 3.33 minutter (e) 3.83 minutter

Kapitel 10

10.1. 0.0078%
10.2 0.18%
10.3 6.1%
10.4 0.24%
10.5 0.024%
10.6 3.04%
10.7 (1) 0.028% (2) -
10.8 25.6%
10.9 3.45%
10.10 (1) 10.6% (2) 1.645 [1.56; 1.73]
10.11 (1) 6.56% (2) 7.7%
10.12 (1) 0.004% (2) 13.1 g [12.8 ; 13.4]
10.13 0.158% [24.07 ; 35.56]
10.14 (1) 4,23% (2) 1049.2 g (3) 0.0314%

10.15 (1) 0.11% (2) [0.69 ; 3.13]
10.16 10.0 %
10.17 3.6%
10.18 0,018% [11.84 ; 15.05]
10.19 7.0%
10.20 (1) 2.39% (2) 0,33%
10.21 4.0%
10.22 15.0%
10.23 (1) 9.8% (2) [0.22 ; 7.28]
10.24 3.63%
10.25 3.73%
10.26 92.47%
10.27 40.29%
10.28 41.6%
10.29 8.77%
10.30 4.26%

Kapitel 11

11.1 (1) - (2) - (3) - (4) - (5) -
11.2 (1) - (2) - (3) - (4) - (5) - (6) -
11.3 (1) - (2) - (3) - (4) - (5) - (6) 1833

Kapitel 12

12.1 (1) - (2) 157.1
12.2 -
12.3 -
12.4 -

STIKORDSREGISTER