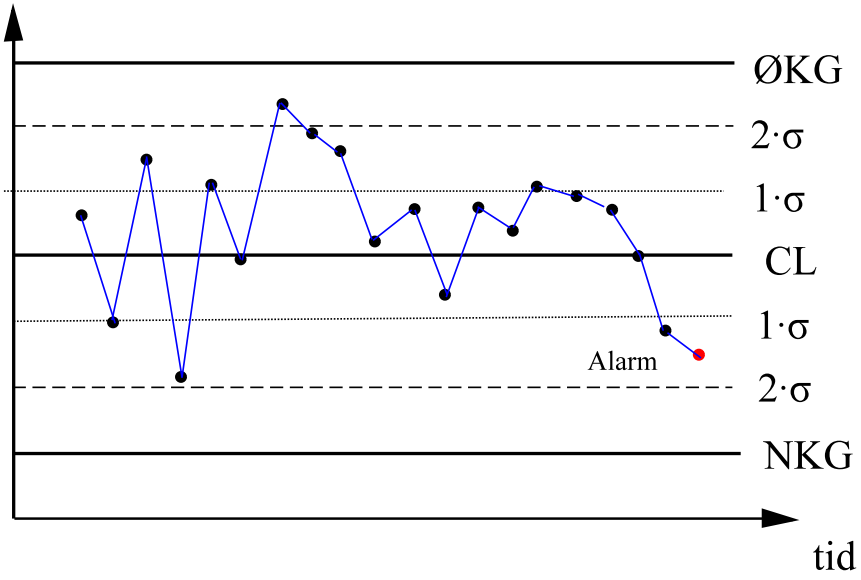
MOGENS ODDERSHEDE LARSEN

Anvendt Statistik

herunder kvalitetskontrol



1. udgave 2018

FORORD

Læsning:
Bogen er bygget således op, at der hurtigt nås frem til normalfordelingen og de vigtige normalfordelingstest. Disse vigtige begreber kan derfor blive grundigt indarbejdet, selv om der kun er kort tid til rådighed.
Er det af tidsmæssige grunde svært at nå hele bogen kan man uden skade for helheden tage de diskrete fordelinger kapitlerne 7 og 8 mere oversigtagtigt
Ligeledes kan emnerne
"Proceskontrol": kapitel 9
"Faktorer på 2 eller flere niveauer": kapitel 12 , 13 og 14
Regressionsanalyse" kapitel 15
læses stort set uafhængigt af hinanden.

Bogen er bygget op således, at de væsentligste begreber søges forklaret anskueligt og ved hjælp af et stort antal eksempler.
I slutningen af nogle af kapitlerne er givet en oversigt over centrale formler eller fremgangs-måder.
Efter hvert kapitel er der nogle opgaver og en facitliste til opgaverne findes bagerst i bogen.
Ønskes en mere dybtgående statistisk forklaring på formlerne kan disse ofte findes på min hjemmeside under titlen "Supplement til anvendt statistik".

Regnemidler
I eksemplerne er beregningerne i videst mulig omfang foretaget ved anvendelse af statistikpro-grammet SAS.JMP.
Ønskes i stedet at anvende TI-Nspire (PC-udgaven) eller det meget udbredte regneark Excel, kan man på hjemmesiden www.larsen-net.dk under statistik 2 hente bøgerne "Statistiske Grundbegreber" og "Videregående statistik", hvor en stor del af bogens eksempler er regnet ved anvendelse af disse programmer.

11. Januar 2018 Mogens Oddershede Larsen

INDHOLD

1 INTRODUKTION TIL STATISTIK 1

2 STOKASTISK VARIABEL
2.1 Sandsynlighed 2
2.2 Stokastisk variabel 3
2.3 Tæthedsfunktion for kontinuert stokastisk variabel 4
2.4 Linearkombination af stokastiske variable 8

3 NORMALFORDELINGEN
3.1 Indledning 10
3.2 Definition og sætninger om normalfordeling 11
3.3 Beregning af sandsynligheder 14
Opgaver 17

4 KONFIDENSINTERVAL FOR NORMALFORDELT VARIABEL
4.1 Udtagning af stikprøver 19
4.2 Fordeling og spredning af gennemsnit 20
4.3 Konfidensinterval for middelværdi 21
4.3.1 Definition af konfidensinterval 21
4.3.2 Populationens spredning kendt eksakt 22
4.3.3 Populationens spredning ikke kendt eksakt 23
4.4 Konfidensinterval for spredning 27
4.5 Oversigt over centrale formler i kapitel 4 30
Opgaver 31

5 HYPOTESETESTNING (1 NORMALFORDELT VARIABEL)
5.1 Grundlæggende begreber 33
5.2 Hypotesetest med ukendt middelværdi og spredning 37
5.3 Fejl af type I og type II 40
5.4 Oversigt over centrale formler i kapitel 5 43
Opgaver 45

6 REGNEREGLER FOR SANDSYNLIGHED, KOMBINATORIK
6.1 Regneregler for sandsynlighed 48
6.2 Betinget sandsynlighed 50
6.3 Kombinatorik 51
6.3.1 Indledning 51
6.3.2 Multiplikationsprincippet 52
6.3.3 Ordnet stikprøveudtagelse 53
6.3.4 Uordnet stikprøveudtagelse 54
Opgaver 55

7 VIGTIGE DISKRETE FORDELINGER
7.1 Indledning 58
7.2 Hypergeometrisk fordeling 58
7.3 Binomialfordeling 60
7.4 Poissonfordeling 66
7.5 Approksimationer 68

7.6 Oversigt over centrale formler i kapitel 7 69
Opgaver 71

8 ANDRE FORDELINGER
8.1 Indledning 76
8.2 Kontinuerte fordelinger 76
8.2.1 Den rektangulære fordeling 76
8.2.2 Eksponentialfordelingen 77
8.2.3 Weibullfordelingen 79
8.2.4 Den logaritmiske normalfordeling 79
8.3 Diskrete fordelinger 80
8.3.1 Den generaliserede hypergeometriske fordeling 80
8.3.2 Polynomialfordelingen 80
Opgaver 82

9 STATISTISK PROCESKONTROL
9.1 Indledning 83
9.2 Statistisk kontrol 84
9.3 Opbygning af kontrolkort 85
9.4 Kontrolkortanalyse 85
9.5 Procesvariablen er normalfordelt 86
9.5.1 Stikprøvestørrelse større end 1 86
9.5.2 Stikprøvestørrelse n = 1 90
9.6 Løbende kontrol 92
9.7 Tolerancegrænser og kapabilitet 93
9.8 Procesvariablen X er diskret 97
4.8.1 X er binomialfordelt 97
4.8.2 X er Poissonfordelt 98
Opgaver 100

10 STATISTISK GODKENDELSESKONTROL
10.1 Indledning 104
10.2 Enkelt stikprøveplan 105
10.3 Rektificerende kontrol 109
10.4 Dobbelt stikprøveplan 109
Opgaver 110

11 PLANLÆGNING AF FORSØG
11.1 Indledning 111
11.2 Nomenklatur 111
11.3 Krav til statistisk gyldigt forsøg 112

12 1 FAKTOR PÅ 2 NIVEAUER
12.1 Normalfordelte variable 116
12.1.1 Test af differens mellem middelværdier 116
12.1.2 Parvise observationer (blokforsøg) 121
12.2 Binomialfordelte variable 122
12.3 Poissonfordelte variable 124
12.4 Fordeling ukendt (rangtest) 124
12.5 Oversigt over centrale formler i kapitel 11 125

12.5.1 Test af differens mellem middelværdier for 2 normal-fordelte variable 125
12.5.2 Test af differens mellem varianser for 2 normalfordelte variable. 127
12.5.3 Test og konfidensintervaller af p_1 - p_2 for 2 binomialfordelte variable. 128
12.5.4 Test og konfidensintervaller af $\mu_1 - \mu_2$ for 2 Poissonfordelte variable 128
Opgaver 129

13 1 FAKTOR PÅ MERE END 2 NIVEAUER
13.1 Indledning 133
13.2 Ensidet variansanalyse (normalfordelte variable) 134
13.2.1 Indledning 134
13.2.2 Forklaring af metoder og formler. 135
13.2.3 Beregning af ensidet variansanalyse 139
13.3 Fuldstændigt randomiseret blokforsøg 142
13.4 Binomialfordelte variable 145
13.5 Poissonfordelte variable. 146
13.6 Oversigt over centrale formler i kapitel 12 147
13.6.1 Oversigt over fremgangsmåde ved ensidet variansanalyse 147
13.6.2 Test af parametre $p_1, p_2, \ldots, p_k$ for binomialfordelte variable. 148
13.6.3 Test af parametre $\mu_1, \mu_2, \ldots, \mu_k$ for Poissonfordelte variable. 148
Opgaver 149

14 2 FAKTORER PÅ 2 ELLER FLERE NIVEAUER. TOSIDET VARIANSANALYSE
14.1 Indledning 152
14.2 Planlægning af forsøg 152
14.2.1 Een faktor ad gangen. 152
14.2.2 Fuldstændig faktorstruktur 153
14.3 Formler og metode. 154
14.4 Beregning af tosidet variansanalyse. 156
14.5 Fuldstændigt randomiseret blokforsøg 163
14.6 Andre varianter af variansanalyse 164
14.6.1 Manglende observationer (ubalancerede forsøg) 164
14.6.2 Trinvis variansanalyse 164
14.8 Oversigt over fremgangsmåde ved tosidet variansanalyse 168
Opgaver 169

15 ENKELT REGRESSIONSANALYSE
15.1 Indledning 173
15.2 Bestemmelse af regressionsligning 174
15.3 Vurdering af om regressionsligning beskriver data godt 176
15.4 Test og konfidensintervaller 182
15.5 Transformation af data 185
15.6 Enkelt regressionsanalyse med flere y - observationer for hver x - værdi 189
15.7 Multipel regressionsanalyse 192
15.7.1 Indledning. 192
15.7.2 Analyse med én y - observation for hver x - værdi 192
15.8 Polynomial regressionsanalyse. 196
15.8.1 Indledning. 196

15.8.2 Beregning af polynomial regressionsanalyse 197
15.9 Oversigt over fremgangsmåde ved regressionsanalyse 202
Opgaver 204

FACITLISTE 211
STIKORD 215

1 INTRODUKTION TIL STATISTIK

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

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

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

I den **deskriptive statistik** beskrives data ved tabeller, grafisk (lagkagediagrammer, søjlediagrammer) og ved beregning af karakteristiske tal såsom gennemsnit og spredning.
Man kan eksempelvis i "Danmarks Statistik" (findes på nettet under adressen www.statistikbanken.dk) finde, hvor mange personbiler der er i Danmark opdelt efter alder.
Man kender her **populationen** (biler i Danmark), kan grafisk vise deres fordeling i et søjlediagram og beregne deres gennemsnitlige alder.
Det forudsættes, at man fra sin tidligere uddannelse kender den diskriptive statistik.
Hvis ikke kan man eksempelvis på min hjemmeside under "statistik 2" , i bogen "Statistiske grundbegreber", kapitel 2 finde en beskrivelse heraf.

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

Eksempelvis udtages ved en meningsmåling en forhåbentlig repræsentativ **stikprøve** på 1000 vælgere, som man spørger om hvilket politisk parti de ville stemme på, hvis der var valg i morgen.
Man vil så ud fra stikprøven konkludere, at hvis man spurgte hele populationen (alle vælgere i Danmark) , så ville man med en vis usikkerhed få samme resultat.
Viser stikprøven, at partiet "Venstre" vil gå 2.5% tilbage, så er det naturligvis usikkert. Man må derfor anvende passende statistiske metoder til eksempelvis at beregne, at dette tal har en usikkerhed på ca 1%.

2 STOKASTISK VARIABEL

2.1 SANDSYNLIGHED

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

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

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

Som eksempler på tilfældige eksperimenter kan nævnes:

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

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

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

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

En hændelse er en delmængde af et eksperiments udfaldsrum.
Eksempelvis er A: "At få et lige øjental" en hændelse ved kast med en terning.

Hændelsen A siges at indtræffe, hvis et udfald fra A forekommer.

Sandsynlighedsbegrebet tager udgangspunkt i begrebet "relativ hyppighed".

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

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

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

Da definitionen af sandsynlighed bygger på relativ hyppighed, er det naturligt, at det for ethvert par af hændelser A og B i udfaldsrummet U skal gælde :
$0 \le P(A) \le 1$, $P(U) = 1$ og P(enten A eller B) = P(A) + P(B) (skrives kort $P(A \cup B) = P(A) + P(B)$) forudsat A og B ingen elementer har fælles (er disjunkte).
(en mere generel regel findes i kapitel 6)
De 3 regler kaldes sandsynlighedsregningens aksiomer.
I kapitel 6 udledes på dette grundlag en række regler for regning med sandsynligheder.

Eksempel 2.1 Regler
Lad A = at få et ulige øjental ved et kast med en terning
B = at få en sekser ved et kast med en terning
Find sandsynligheden for enten at få et ulige øjental eller en sekser (evt. begge dele) ved kast med en terning.
Løsning:

$P(A) = \frac{1}{2}$. $P(B) = \frac{1}{6}$ $\qquad P(A \cup B) = P(A) + P(B) = \frac{1}{2} + \frac{1}{6} = \frac{2}{3}$ ◆

2.2 STOKASTISK VARIABEL

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

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

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

Vi vil i senere afsnit behandle diskrete variable.

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

2.3 TÆTHEDSFUNKTION FOR KONTINUERT STOKASTISK VARIABEL

Eksempel 2.2. Histogram for kontinuert stokastisk variabel

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

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

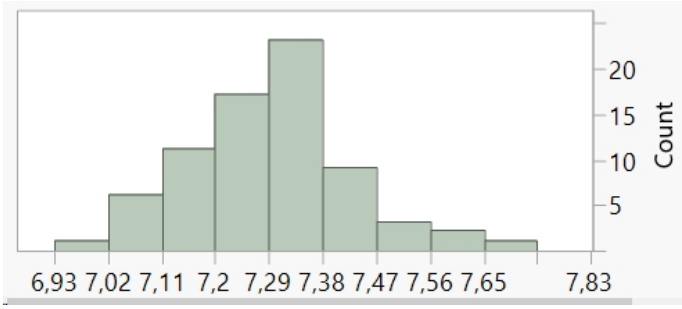
Population og stikprøve. Samtlige indbyggere i Norden med denne sygdom udgør **populationen**. Da det er ganske uoverkommeligt at undersøge alle, udtages en **stikprøve** på 75 patienter.

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

Histogram. For at få et overblik over et større datamateriale, vil man sædvanligvis starte med at tegne et histogram.
Først findes det største tal x_{max} og det mindste tal x_{min} i materialet og derefter beregne **variationsbredden** $x_{max} - x_{min}$.
Vi finder, at største tal er 7.71 og mindste tal er 6.95.
Variationsbredden er derfor 7.71 - 6.95 = 0.76.

Dernæst deles tallene op i et passende antal intervaller (klasser).
Som det første bud vælges ofte et antal nær $\sqrt{n}$. Da $\sqrt{75} \approx 9$ vælges ca. 9 klasser. Da $\frac{0.76}{9} \approx 0.09$ vælges det som intervalbredde.

SAS.JMP
Analyze ▶ Distribution ▶ Vælg “Histograms Only og marker pH og indsæt i Y,Columns ▶ ok
rød pil ved pH ▶ Histogram Options ▶ stryg Vertical og “set Bin With “ til 0.09 og vælg “Count Axis



Histogrammet er et "klokkeformet histogram", hvor der på tegningen er flest tal fra 7.29 til 7.29+0.09=7.38 og derefter falder antallet til begge sider. ◆

Tæthedsfunktion

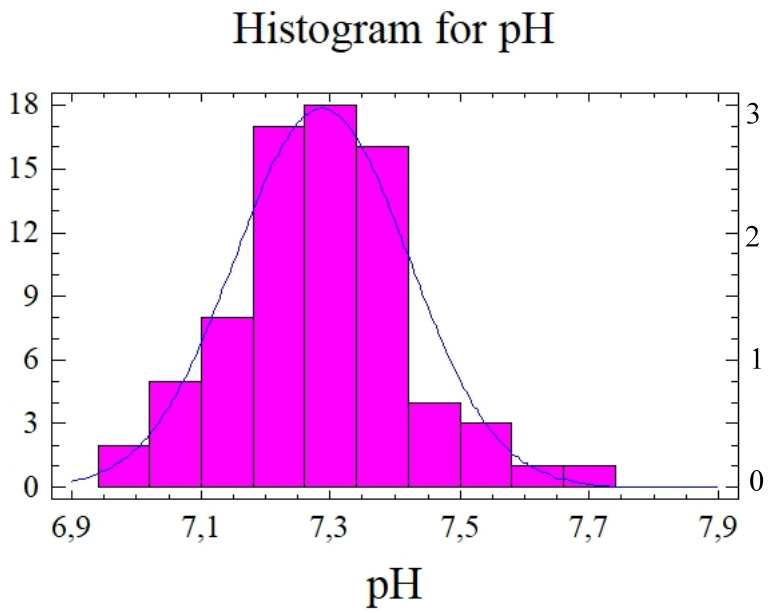
Vi fortsætter med eksempel 2.2

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).

Endvidere er der beregnet en søjle ved at dividere den relative hyppighed med intervallængden.

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

Allerede her kan man se, at antallet er størst omkring 7.30, 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. Hvis man tænker sig antallet af forsøg stiger (for eksempel undersøger hele populationen på måske 1 million nordiske knæ), samtidig med at man øger antallet af klasser tilsvarende (til for eksempel $\sqrt{10^6} \approx 1000$) , vil histogrammet blive mere og mere fintakket, og til sidst nærme sig til en kontinuert klokkeformet kurve (indtegnet på grafen).

Hvis man benytter den skalerede skala fra skemaet, som også er afsat på højre side af tegningen, vil arealet af hver søjle være den relative hyppighed, og for den idealiserede kontinuerte kurve, vil arealet under kurven i et bestemt interval fra a til b være sandsynligheden for at få en værdi mellem a og b. Det samlede areal under kurven er naturligvis 1. ◆

Man siger, at den kontinuerte stokastiske variabel X (p værdien) har en **tæthedsfunktion f(x)** hvis graf er den ovenfor nævnte kontinuerte kurve.
Da arealet under en kontinuert kurve beregnes ved et bestemt integral, følger heraf følgende definition:

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

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

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

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

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

Måleresultater vil sædvanligvis være værdier af normalfordelte variable, så en rimelig hypotese for den i eksempel 2.2 angivne kontinuerte stokastiske variabel X = p er således, at den er normalfordelt. Dette bestyrkes af at grafen for sådanne netop er klokkeformede .

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

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

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

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

Eksempel 2.3 Kontinuert stokastisk variabel.

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

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

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

b) Skitser grafen for f.

c) Beregn middelværdi og spredning for X.

Løs

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

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

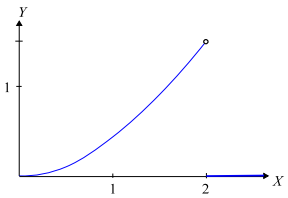


Fig.3.1 *Tæthedsfunktion*

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

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

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

DEFINITION af fordelingsfunktion F(x) for kontinuert variabel.

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

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

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

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

Eksempel 2.4. Fordelingsfunktion for kontinuert variabel.
For den i eksempel 2.3 angivne kontinuerte variabel X med tæthedsfunktion $f(x)$ ønskes fundet:
1) Fordelings[illegible]
2) Medianen
Løsning:

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

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

2.4 LINEARKOMBINATION AF STOKASTISKE VARIABLE

Vi betragter i dette afsnit flere stokastiske variable.
Eksempel 2.5 vil blive benyttet som gennemgående eksempel

Eksempel 2.5. To variable.
Insektpulver sælges i papkartoner. Lad den stokastiske variable X_1 være vægten af pulveret, mens X_2 er vægten af papkartonen. I middel fyldes der 500 gram insektpulver i hver karton med en spredning på 5 gram. Kartonen vejer i middel 10 gram med en spredning på 1.0 gram.
$Y = X_1 + X_2$ er da bruttovægten.
1) Find middelværdien af Y
2) Find spredning af Y.

◆

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

Man kan vise at der gælder følgende
Linearitetsregel: $E(Y) = a_0 + a_1 \cdot E(X_1) + a_2 \cdot E(X_2) + \ldots + a_n \cdot E(X_n)$.

I eksempel 2.5 synes det rimeligt at antage, at vægten af pulveret og vægten af papkartonen er **uafhængige** (påfyldningen kan tænkes at ske maskinelt, uden at den er afhængig på nogen måde af hvilken vægt, kartonen tilfældigvis har).

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

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

Eksempel 2.5. (fortsat) **To variable.**
Spørgsmål 1: $E(Y) = E(X_1) + E(X_2) = 500 + 10 = \underline{510 \text{ gram}}$.
Spørgsmål 2: $V(Y) = V(X_1) + V(X_2) = 5^2 + 1^2 = 26$. $\sigma(Y) = \sqrt{26} = \underline{5.1}$ gram. ◆

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

Eksempel 2.6. Ensfordelte variable
Bruttovægten af det i eksempel 2.5 nævnte karton insektpulver havde middelvægten 510 g med en spredning på 5.1 g.
Vi udtager nu tilfældigt og uafhængigt af hinanden 10 pakker insektpulver.
a) Hvad bliver i middel den samlede vægt af de 10 kartonner
b) Hvad bliver i middel spredningen på den samlede vægt af de 10 kartoner
Løsning:
Lad X_1 være vægten af karton 1, X_2 være vægten af karton 2 osv. X_{10} være vægten af karton 10.
$Y= X_1 + X_2 + ... + X_{10}$ er da vægten af alle 10 kartonner.
a) $E(Y) = E(X_1)+E(X_2)+ \ldots +E(X_{10}) =10 \cdot 510 = \underline{\underline{5100 \text{ g}}}$
b) $V(Y) = V(X_1)+V(X_2)+ \ldots +V(X_{10}) =10 \cdot (5.1)^2 = 260.1$ g

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

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

For ensfordelte uafhængige stokastiske variable gælder:

SÆTNING 2.1 (*middelværdi og spredning for stikprøves gennemsnit*)

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

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

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

Eksempel 2.7. Spredning på gennemsnit (eksempel 2.5 fortsat)
Hvis der udtages 5 kartoner insektpulver, hvad vil da være spredningen på gennemsnittet af vægten af insektpulveret .
Løsning:
Da spredningen på 1 karton er 5.1 gram, vil spredningen på gennemsnittet af 5 kartoner være

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

Opgave 2.1
Vægten af en (tilfældigt udvalgt) tablet af en vis type imod hovedpine har middelværdien $\mu = 0.65$ g og spredningen $\sigma = 0.04$ g
a) Beregn middelværdi og spredning af den sammenlagte vægt af 100 (tilfældigt udvalgte) tabletter
b) På basis af de 100 tabletter ønskes spredningen på gennemsnittet beregnet.

3 NORMALFORDELINGEN

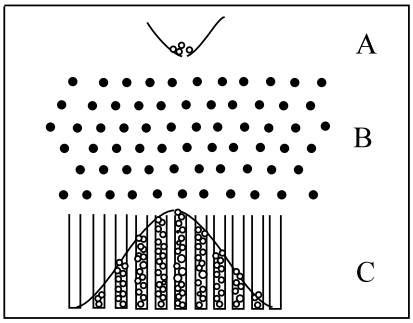
3.1 INDLEDNING

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

Som illustration af dette kan anvendes Galtons apparat.

Eksempel 3.1. Eksperiment med et Galton-apparat.

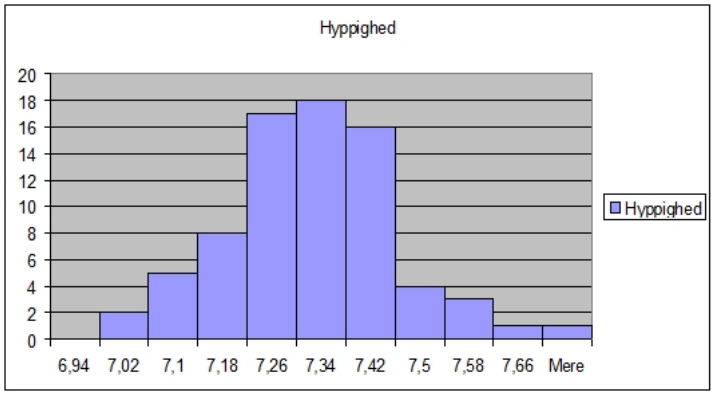
På den anførte figur er skitseret et Galton-apparat.
A er en tragt; B er sømrækker, hvor sømmene i en underliggende række er anbragt midt ud for mellemrummene mellem sømmene i den overliggende række; C er opsamlingskanaler.



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

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

En anden illustration af under hvilke omstændigheder en normalfordelt variabel kan forekomme i praksis så vi i kapitel 2 eksempel 2.2 hvor man på 75 mennesker med en bestemt ledsygdom målte pH i knæleddet.
Histogrammet som er gentaget nedenfor har et klokkeformet udseende, som kraftigt antyder, at den kontinuerte stokastiske variabel X = pH er normalfordelt.



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

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

3.2 DEFINITION OG SÆTNINGER OM NORMALFORDELING

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

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

Den har middelværdi μ og spredning σ

Grafen er klokkeformet og symmetrisk om linien x = μ.

◆

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

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

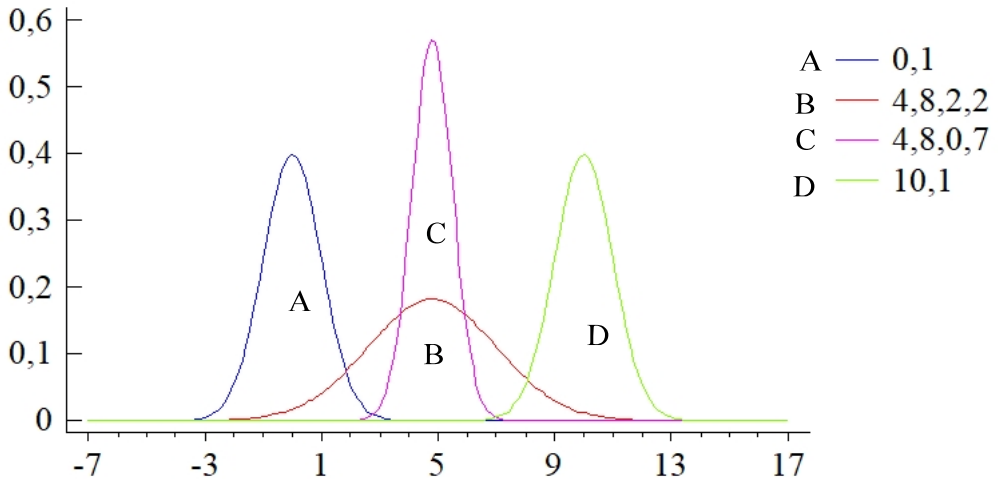


Fig 3.1 *Forskellige normalfordelinger*

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

Vi nævner uden bevis følgende sætning:

SÆTNING 3.1 Additionssætning for linearkombination af normalfordelte variable.
Er Y en linearkombination af n stokastisk uafhængige , normalfordelte variable, vil Y også være normalfordelt.

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

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

Normeret normalfordeling

Af særlig interesse er den såkaldte normerede normalfordeling.
Den er bestemt ved at have middelværdien 0 og spredningen 1.
Grafen for den er tegnet som graf A i figur3.1
Den kaldes sædvanligvis *U* eller Z og dens fordeling **U- eller Z-fordelingen** . Dens tæthedsfunktion benævnes φ og dens fordelingsfunktion Φ.
Specielt vil dens p - fraktil z_p indgå i adskillige formler i de næste afsnit.

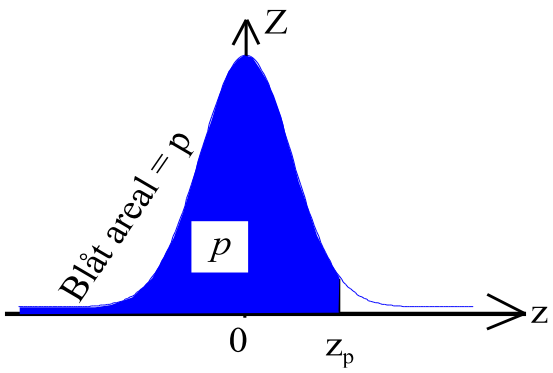


Fig3.2 *Normeret normalfordeling*

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

(3.1) $x_p = z_p \cdot \sigma + \mu$ eller omvendt $z = \dfrac{x-\mu}{\sigma}$

0 Beviset for denne relation indgår i beviset for den følgende sætning, som også viser, at man kan overføre en vilkårlig normalfordeling til den normerede normalfordeling.
Det er derfor nok at kunne lave et program, der kan foretage beregninger i den normerede normalfordeling.
Der gælder følgende

SÆTNING 3.2. (normering af normalfordeling). *Når X er normalfordelt* $n(\mu,\sigma)$

er den variable $Z = \dfrac{X-\mu}{\sigma}$ *normalfordelt* $n(0,1)$, *og der gælder*

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

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

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

Bevis:
At *Z* også er normalfordelt vises ikke her.

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

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

Z har derfor middelværdi 0 og spredning 1.

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

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

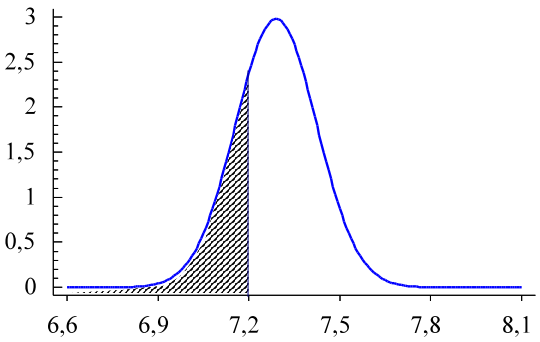
Bevis for $x_p = z_p \cdot \sigma + \mu$: $P(X \le x_p) = p \Leftrightarrow \Phi\left(\frac{x_p-\mu}{\sigma}\right) = p \Leftrightarrow \frac{x_p-\mu}{\sigma} = z_p \Leftrightarrow x_p = z_p \cdot \sigma + \mu$

3.3. BEREGNING AF SANDSYNLIGHEDER

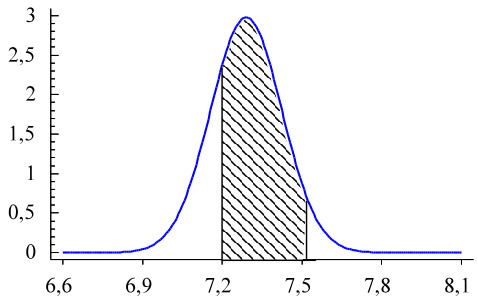
Stikprøves gennemsnit og spredning.

Ofte er middelværdien μ og spredningen σ ukendt i en foreliggende normalfordeling. I så fald erstattes fordelingen $n(\mu,\sigma)$ i praksis med en approksimerende fordeling $n(\bar{x},s)$, såfremt der foreligger et rimelig stort antal observationer fra den givne fordeling.
På basis af den i eksempel 2.5 angivne stikprøve på 75 patienter beregnes et gennemsnit af pH værdierne på $\bar{x} = \frac{546.52}{75} = 7.2868$ og en s værdi på $s = \sqrt{\frac{SAK}{n-1}} = 0.134355$.
Vi vil altså antage, at pH værdierne er approksimativt normalfordelt n (7.29, 0.134).

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



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



Eksempel 3.2. Beregning af normalfordelte sandsynligheder

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

1) Find $P(X \le 7.2)$
2) Find $P(7.2 \le X \le 7.5)$
3) Find $P(X > 7.6)$
4) Find $x_{0.9}$ bestemt ved, at $P(X < x_{0.9}) = 0.9$

Løsning:

File ▶ New ▶ Data Table ▶ Giv en søjle en overskrift ▶ Skriv et tilfældigt tal f.eks 1 i kolonnen nedenunder ▶
Cursor på overskriften, højre musetast ▶ Vælg i menu "Formula" ▶ vælg Probability ▶ Normal Distribution skriv 7,2 ▶ sæt komma, hvorved der nu står Mean, skriv 7,29 ▶ igen komma hvorved der nu står std dev, skriv 0,134 ▶ ok

1) $P(X \le 7.2) =$ Normal Distribution(7,2, 7,29, 0,134) $= 0.2509$

2) $P(7.2 \le X \le 7.5) = P(X \le 7.5) - P(X<7.2) =$ Normal Distribution(7,5, 7,29, 0,134) - Normal Distribution(7,2, 7,29, 0,134)
$= 0.6906$

Husk at bruge JMP-tastaturet når man skal skrive minus.

3) $P(X>7.6) = 1 - P(X \le 7.6) =$ 1 - Normal Distribution(7,6, 7,29, 0,134) $= 0.0103$

4) Har man omvendt givet en sandsynlighed $p = 0.9$ og ønsker at finde den tilsvarende værdi x_p for hvilken $P(X \le x_p) = 0.9$ betyder det, at man kender arealet 0.9 og skal finde x-værdien. Det svarer jo til at finde den inverse (omvendte) funktion af normalfordelingen.

$x_{0.9} =$ Normal Quantile(0,9, 7,29, 0,134) $= 7.46$ ◆

Eksempel 3.3. Kvalitetskontrol.

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

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

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

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

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

d) Find de nye tolerancegrænser (placeret symmetrisk omkring middelværdien 90,0) idet spredningen stadig er 0.5, og højst 5% må kasseres.

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

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

Løsning:
Vælg i søjle Formula

a) $P(89.2 \le X \le 90.8)$= Normal Distribution(90,8, 89,6, 0,5) - Normal Distribution(89,2, 89,6, 0,5) = 0.7799 = 77.99%

b) Middelværdien justeres til midtpunktet 90.0.

P(X>90.8)+P(X<89.2)= 1-P(X<90.8) + P(X<89,2) =

(1 - Normal Distribution(90,8, 90, 0,5)) + Normal Distribution(89,2, 90, 0,5) =0.1096= 10.96%

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

$$89.2 = z_{0.025} \cdot \sigma + 90 \Leftrightarrow \sigma = \frac{-0.8}{z_{0.025}} \Leftrightarrow \sigma =$$ (0 - 0,8) / Normal Quantile(0,025,) =0.4082

d) Kaldes den nedre tolerancegrænse for a fås med samme begrundelse som i punkt c : $P(X \le a) = 0.025$. Vi kan her benytte den "inverse" normalfordeling

Nedre grænse a = Normal Quantile(0,025, 90, 0,5) = 89.02

Øvre grænse b = 90 +(90 - 89.02) = 90.98

e) Data indtastes i kolonne med navn x ▶ Analyze ▶ Distribution ▶ x indsættes i Y,columns ▶ OK, ▶ nederst trykkes på rød pil og man vælger "Customize Summery Statistics" ▶ Marker Mean og Stdev (slet resten)
Gennemsnit: 89.97 Spredning: 0.4355

Summary Statistics	
Mean	89,966667
Std Dev	0,4355422

Eksempel 3.4. Additionssætning.
En boreproces fremstiller huller med en diameter X_1, der er normalfordelt med en middelværdi μ_1=10 og en spredning på 0.04. En anden proces fremstiller aksler med en diameter X_2, der er normalfordelt med en middelværdi $\mu_2 = 9.94$ 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. $E(Y) = E(X_2) - E(X_1) = 9.94 - 10.00 = -0.06$
$V(Y) = 1^2 V(X_2) + (-1)^2 V(X_1) = 0.04^2 + 0.03^2 = 0.0025 \quad \sigma(Y) = \sqrt{0.0025} = 0.05$

$P(X_2 < X_1)$= P(Y<0) = Normal Distribution(0, 0 - 0,06, 0,05) = 0.8849 = 88.49% ◆

OPGAVER

Opgave 3.1

1) En stokastisk variabel X er normalfordelt med $\mu = 23.1$ og $\sigma = 2.4$.
   Find $P(X \le 22.2)$, $P(X > 25.3)$ og $P(22.3 < X \le 27.8)$.
2) Find tallet x, hvor P(X<x) = 0.1.

Opgave 3.2

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

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

Man overvejer at gå over til en anden opvarmningsproces. Man udfører derfor 16 gange i løbet af en periode forsøg, hvor man måler maksimumstemperaturen, der opnås ved denne nye proces. Resultaterne var

116.6 , 116,6 , 117,0 , 124,5 , 122,2 , 128,6 , 109,9 , 114,8 , 106,4 , 110,7, 110,7 , 113,7 , 128,1, 118,8 , 115,4 , 123,1

4) Giv et estimat for middelværdien og spredningen.

Opgave 3.3

En fabrik planlægger at starte en produktion af rør, hvis diametre skal opfylde specifikationerne 2,500 cm ± 0,015 cm.

Ud fra erfaringer med tilsvarende produktioner vides, at de producerede rør vil have diametre, der er normalfordelte med en middelværdi på 2,500 cm og en spredning på 0,010 cm. Man ønsker i forbindelse med planlægningen svar på følgende spørgsmål:

1) Hvor stor en del af produktionen holder sig indenfor specifikationsgrænserne.
2) Hvor meget skal spredningen σ ned på, for, at 95% af produktionen holder sig indenfor specifikationsgrænserne (middelværdien er uændret på 2,500 cm).
3) Fabrikken overvejer, om det er muligt at få indført nogle specifikationsgrænser (symmetrisk omkring 2,500), som bevirker, at 95% af dets produktion falder indenfor grænserne. Find disse grænser, idet det stadig antages at middelværdien er 2.500 og spredningen 0.010 cm.

Opgave 3.4

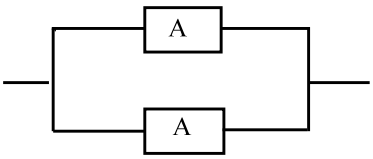
En automatisk dåsepåfyldningsmaskine fylder hønskødssuppe i dåser. Rumfanget er normalfordelt med en middelværdi på 800 ml og en spredning på 6,4 ml.

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

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

Opgave 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.
Beregn sandsynligheden p for, at en tilfældig pakke opvaskemiddel har en bruttovægt mellem 4250 g og 4300 g.

Opgave 3.7
Et system er af sikkerhedsmæssige grunde opbygget af to apparater A, der er parallelforbundne (se figur) således, at systemet virker, så længe blot et af apparaterne virker.



Svigter et af apparaterne, startes reparation. Det antages, at reparationstiden er normalfordelt med middelværdien
μ_{rep} = 10 timer og spredning σ_{rep} = 3 timer.
I reparationstiden overbelastes den anden komponent, og det antages, at dens levetid fra reparationens start (approksimativt) er normalfordelt med middelværdi μ og spredning σ = 4 timer.
1) Find sandsynligheden for, at reparationen er afsluttet, inden den anden komponent fejler, hvis μ = 20 timer.
2) Hvor stor skal μ være, for at sandsynligheden for, at reparationen kan afsluttes før den anden komponent fejler, er mere end 99.9%?

Opgave 3.8
Vægten af en (tilfældig udvalgt) tablet af en vis type mod hovedpine har middelværdien 0.65 g og spredningen 0.04 g.
1) Beregn middelværdi og spredning af den sammenlagte vægt af 100 (tilfældigt udvalgte) tabletter.
2) Antag, at man benytter følgende metode til at fylde tabletter i et glas. Man placerer glasset på en vægt og fylder tabletter på, indtil vægten af tabletterne i glasset overstiger 65,3 g. Beregn sandsynligheden for, at glasset kommer til at indeholde mere end 100 tabletter (se bort fra vægtens fejlvisning).

4. Konfidensinterval for normalfordelt variabel

4.1 UDTAGNING AF STIKPRØVER

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

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

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

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

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

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

4.2. FORDELING OG SPREDNING AF GENNEMSNIT

Udtages en stikprøve fra en population er det jo for, at man ud fra stikprøven kan fortælle noget centralt om hele populationen.
I eksempel 2.2 var vi således interesseret i koncentrationen af brintioner (pH) i ledvæsken i knæet hos patienter, der led af denne sygdom.
Som led i en nordisk medicinsk undersøgelse udtog man blandt patienter der led af denne sygdom tilfældigt en stikprøve på 75.
På basis heraf beregnede man gennemsnittet af pH værdierne til $\bar{x}$ = 7.2868 og spredningen s = 0.134355 .
Man vil nu sige, at et estimat (skøn) for den "sande" middelværdi μ for hele populationen er 7.29 og den "sande" spredning" σ er 0.134.

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

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

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

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

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

Eksempel 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 $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{50}}$.

Vi har derfor

$P(7 < \overline{X} < 9) = P(\overline{X} < 9) - P(\overline{X} < 7)$ = Normal Distribution(9, 8,2, $\frac{3}{\sqrt{50}}$) - Normal Distribution(7, 8,2, $\frac{3}{\sqrt{50}}$)

=0.968 = 96.8%

4.3. KONFIDENSINTERVAL FOR MIDDELVÆRDI

4.3.1 Definition af konfidensinterval

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

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

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

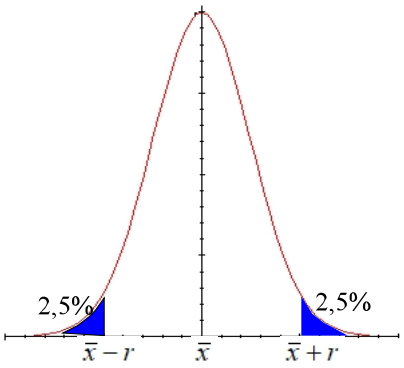
Det er derfor ikke nok, at angive at den "sande" middelværdi er $\bar{x}$, vi må også angive et "usikkerhedsinterval".

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

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

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

4.3.2. Populationens spredning kendt eksakt



Et 95% konfidensinterval $[\bar{x}-r;\bar{x}+r]$ må ligge symmetrisk omkring gennemsnittet, og således, at
$P(\bar{x}-r \le \bar{X} \le \bar{x}+r) = 0.95$.
Heraf følger, at hvis den sande middelværdi μ ligger i et af de farvede områder på figuren, så er der mindre end 2.5% chance for, at vi ville have fået det fundne gennemsnit $\bar{x}$.

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

Man må her huske, at et gennemsnit har spredningen $\frac{\sigma}{\sqrt{n}}$, hvor σ er spredningen på den enkelte måling og *n* er antal målinger i stikprøven.
Fremfor at løse ovenstående ligning, er det lettere at benytte formlen i sætning 4.1

Sætning 4.1. Spredning kendt eksakt
Er spredningen eksakt kendt er et 95% konfidensinterval bestemt ved formlen

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

Bevis:

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

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

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

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

Eksempel 4.2. Konfidensinterval hvis spredningen er kendt eksakt

Lad gennemsnittet af 12 målinger være $\bar{x} = 90$, og lad os antage, at spredningen kendes eksakt til $\sigma = 0.5$.

Bestem et 95% konfidensinterval for middelværdien μ.

Løsning:

Radius r = $\text{Normal Quantile}(0{,}975) \cdot \left(\frac{0{,}5}{\sqrt[2]{12}}\right)$

Nedre grænse n: P(X<n)=0.025 n = 90 -r =89,717

Øvre grænse ø: PX<ø)=0.975 ø = 90 + r = 90.283

95% konfidensinterval: [89.717 ; 90.283]

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

4.3.3. Populationens spredning ikke kendt eksakt

Sædvanligvis er populationens spredning σ jo ikke eksakt kendt, men man regner et estimat *s* ud for den.

Da *s* jo også varierer fra stikprøve til stikprøve, giver dette en ekstra usikkerhed, så konfidensintervallet for μ bliver bredere.

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

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

t-fordelinger

En t - fordeling har samme klokkeformede udseende som en Z - fordeling (en normalfordeling med middelværdi 0 og spredning 1)

I modsætning til Z - fordelingen afhænger dens udseende imidlertid af antallet *n* af tal i stikprøven.

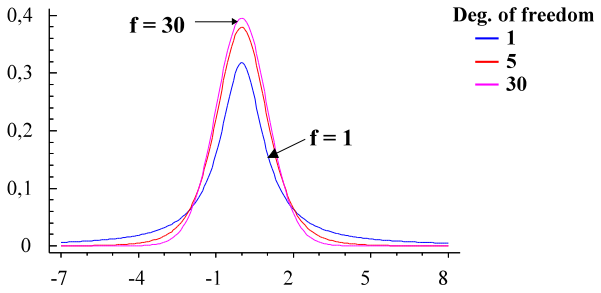
Er **frihedsgradstallet** $f = n$ -1 stort (over 30) er forskellen mellem en U- fordeling og en t-fordeling så lille, at den sædvanligvis er uden praktisk betydning.

Er f lille bliver t - fordelingen så meget bredere end Z - fordelingen, at t-fordelingen må anvendes i stedet for Z-fordelingen.

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

---

[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.



Eksempel 4.3. Beregning af t-værdier.

1) Find $t_{0.975}(12)$ og $t_{0.025}(12)$.

2) Find $P(X \geq 1)$, hvor X er t - fordelt med 12 frihedsgrader.

Løsning:

Formula ▶ Statistik ▶ Probability ▶ tQuantile ▶ indsæt data

1) $t_{0.975}(12)$ = tQuantile(0,975,12) = 2.18 $t_{0.025}(12)$ = tQuantile(0,025,12) = -2.18

2) $P(X \geq 1)$ = 1 - tDistribution(1, 12) = 0.1685 = 16.85%

◆

Er den eksakte værdi af spredningen ukendt, erstattes den i den af sætning 4.1 angivne formel med spredningen s.
Idet s har frihedsgradstallet n - 1 erstattes $z_{0.975}$ med $t_{0.975}(n - 1)$
Herved fremkommer sætningen.

Sætning 4.2. (Den eksakte værdi af spredning ukendt)

Et 95 % konfidensinterval er bestemt ved formlen: $\bar{x} - t_{0,975}(n-1) \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{0.975}(n-1) \cdot \frac{s}{\sqrt{n}}$

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

Eksempel 4.4 :Konfidensinterval, hvor originale data ikke er kendt.

Find konfidensintervallet for middelværdien μ, idet stikprøven er på 12 tal, som man fundet giver et gennemsnit på 90 og et estimat for spredningen på 0.5. (samme data som i eksempel 4.2)

Løsning:

Man benytter formlen i sætning 4.2

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

Radius r = $\text{tQuantile}(0{,}975, 11) \cdot \left(\frac{0{,}5}{\sqrt[n]{12}}\right)$ = 0.3177

Nedre grænse :90 - r =89.68 Øvre grænse: 90 + r = 90.32

95% konfidensinterval :[89.68 ; 90,32]

Sammenlignes med eksempel 4.2 ses som forventet, at konfidensintervallet er bredere. ◆

Eksempel 4.5. Konfidensinterval, hvor originale data er kendt.
Ved fremstilling af et bestemt levnedsmiddel er det vigtigt, at et tilsætningsstof findes i levnedsmidlet i en koncentration på 8.50 (g/l).
For at kontrollere dette udtager levnedsmiddelkontrollen 6 prøver af levnedsmidlet. Resultaterne var:

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

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

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

Løsning
SAS.JMP har indbygget et program, så man ikke behøver at anvende formlerne direkte.
Data indsættes i søjle med navn x

x
8,54
7,89
8,5
8,21
8,15
8,32

a) Analyze ▶ Distribution ▶ x indsættes i Y,Columns ▶ OK, ▶ nederst fremkommer følgende tabel

Summary Statistics	
Mean	8,2683333
Std Dev	0,2411155
Std Err Mean	0,098435
Upper 95% Mean	8,5213685
Lower 95% Mean	8,0152982
N	6

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

b) 95% konfidensinterval: [8.02 ; 8.52]

Da intervallet indeholder 8.50, er kravet opfyldt, men da intervallet kun lige netop indeholder tallet 8.50, så det vil nok være rimeligt, at foretage en ny vurdering på basis af nogle flere målinger. ♦

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

SÆTNING 4.2 ($100\cdot(1-\alpha)$ *% prædistinationsinterval for en enkelt observation)*.
Et $100\cdot(1-\alpha)$ *% prædistinationsinterval for en enkelt fremtidig observation* X_{n+1} *er bestemt ved*

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

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

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

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

Eksempel 4.6. Prædistinations-interval for middelværdi af normalfordeling.

Samme problem som i eksempel 4.5, men nu ønskes bestemt et 95% prædistinationsinterval for en enkelt ny måling af koncentrationen.

Løsning

Da konfidensintervallet har længden 8.52 - 8.02 = 0.50 er radius $r_k = 0.25$

Vi har derfor $r_p = 0.25\cdot\sqrt{6+1} = 0.66$ og dermed

95% prædistinationsinterval $= [8.27 - 0.66;\ 8.27 + 0.66] == \underline{\underline{[7.61\ ;\ 8.93]}}$.

SAS.JMP: Rød pil ved x ▶ Prediction Interval ▶ skriv 0,95 og vælg Two-sided ▶ OK

Prediction Interval

Parameter	Future N	Lower PI	Upper PI	1-Alpha
Individual	1	7,598865	8,937801	0,950
Mean	1	7,598865	8,937801	0,950
Std Dev	1	.	.	0,950

Bestemmelse af stikprøvens størrelse

Før man starter sine målinger, kunne det være nyttigt på forhånd at vide nogenlunde hvor mange målinger man skal foretage, for at få resultat med en given nøjagtighed.

Hvis spredningen antages kendt , ved vi, at radius i konfidensintevallet er $r = z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}$

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

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

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

Det følgende eksempel illustrerer fremgangsmåden.

Eksempel 4.7. Bestemmelse af stikprøvens størrelse.

En forstmand er interesseret i at bestemme middelværdien af diameteren af voksne egetræer i en bestemt fredet skov.
Der blev målt diameteren på 7 tilfældigt udvalgte egetræer (i 1 meters højde over jorden)
På basis af målingerne på de 7 træer sættes $s \approx 14$.

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

Løsning:

a) Først benyttes formlen $n = \left(\frac{z_{0.975} \cdot s}{r} \right)^2$

$\left(\frac{\text{Normal Quantile}(0{,}975) * 14}{5} \right)^2 = 30{,}12 \approx 31$

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

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

Da $n < 30$ burde man have anvendt en t - fordeling. $n = \left(\frac{t_{0.975,}(n-1) \cdot s}{r} \right)^2$

2 søjler benævnes n og n1
I søjle n skrives 22 og i søjle n1 skrives
Derefter skrives 23 i søjle n osv.

$\left(\frac{\text{t Quantile}(0{,}975, n-1) * 14}{6} \right)^2$

Resultat

n	n1
22	23,546099268
23	23,416280621
24	23,29865235

Heraf ses, at n= 24, da vi jo må vælge den største værdi

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

I visse situationer ønsker man at finde et konfidensinterval for spredningen.
Vi vil ikke gå nærmere ind på teorien herfor, men blot henvise til formlerne i oversigt 4.5.

Er den eksakte værdi af middelværdien ukendt skal følgende formel benyttes: $\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}$

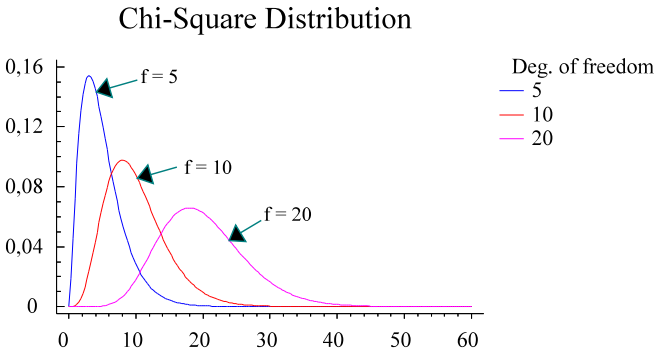
I formlerne indgår den såkaldte χ^2 - fordeling, (udtales ki i anden) .

χ^2 -fordelinger

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

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

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



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

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

Eksempel 4.8. Beregning af χ^2 - værdier.

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

Løsning:

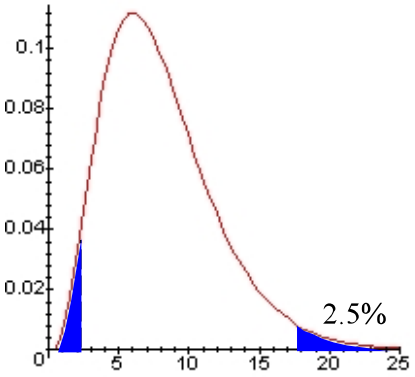
Formula, Probability, ChiSquare Quantiler, udfyld menu

1)

$\chi^2_{0.025}(8)$ = ChiSquare Quantile(0,025, 8) = 2.1797

$\chi^2_{0.975}(8)$ = ChiSquare Quantile(0,975, 8) = 17.5345

2) $P(X \le 5)$ = ChiSquare Distribution(5, 8) = 0.2424



Eksempel 4.9. Konfidensinterval for varians og spredning af normalfordeling.

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

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

a) Angiv på basis af måleresultaterne et estimat for opløsningens spredning.
b) Angiv et 95% konfidensinterval for spredningen.

Løsning:

a) Data indtastes i kolonne med navn x ▶ Analyze ▶ Distribution ▶ x indsættes i Y,Columns ▶ OK, ▶ nederst trykkes på rød pil og man vælger "Customize Summery Statistics" ▶ Marker Stdev og Variance (slet resten)

Resultat findes nederst:

Spredning s = 0.3162

Summary Statistics	
Std Dev	0,3162278
Variance	0,1

b) Der må anvendes formel 3 i oversigt 4.5 : $\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}$

Formula: Nedre grænse for vari- $\frac{(11 * 0{,}1)}{\text{ChiSquare Quantile}(0{,}975,\ 11)}$ ans: =0,05018

Øvre grænse for varians $\frac{(11 * 0{,}1)}{\text{ChiSquare Quantile}(0{,}025,\ 11)}$ = 0.2883

95% konfidensinterval for varians [0.05018 ; 0.2883]

95% konfidensinterval for spredning : [0.2240 ; 0.5369]

hvor $\sqrt{\text{ned-varians}}$ = 0.2240

4.5. OVERSIGT over centrale formler i kapitel 4

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

Oversigt over konfidensintervaller

nr	Forudsætninger	Estimat for parameter	100 (1 - α) % konfidensinterval for parameter
1	μ ukendt. σ ukendt	For μ: $\bar{x}$	$\bar{x} - t_{1-\frac{\alpha}{2}}(n-1)\cdot\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{1-\frac{\alpha}{2}}(n-1)\cdot\frac{s}{\sqrt{n}}$ Metode: se eksempel 4.4 eller eksempel 4.5
	μ ukendt. σ kendt	For μ: $\bar{x}$	$\bar{x} - z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}$ Metode : se eksempel 4.2
3	μ ukendt σ ukendt.	For σ^2: s^2	$\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}$ Metode: se eksempel 4.9
4	μ kendt σ ukendt.	For σ^2: $s_\mu^2 = \frac{(n-1)s^2 + n(\bar{x}-\mu)^2}{n}$	$\frac{(n-1)s^2 + n(\bar{x}-\mu)^2}{\chi^2_{1-\frac{\alpha}{2}}(n)} \le \sigma^2 \le \frac{(n-1)s^2 + n(\bar{x}-\mu)^2}{\chi^2_{\frac{\alpha}{2}}(n)}$

Oversigt over prædistinationsintervaller

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

Bestemmelse af stikprøvens størrelse *n*.

Ønsket værdi af radius *r* i 100 (1 - α) % konfidensinterval			
1	σ kendt eller n > 30	$n = \left(\frac{z_{1-\frac{\alpha}{2}}\cdot\sigma}{r}\right)^2$	Metode: se eksempel 4.7
2	σ ukendt, men antag den højst er s	$n = \left(\frac{t_{1-\frac{\alpha}{2}}(n-1)\cdot s}{r}\right)^2$	Metode: se eksempel 4.7

OPGAVER

Opgave 4.1

En fabrik producerer stempelringe til en bilmotor. Det vides, at stempelringenes diameter er approksimativt normalfordelt. Stempelringene bør have en diameter på 74.036 mm.
For at kontrollere dette udtog man tilfældigt 15 stempelringe af produktionen og målte diameteren. Man fandt, at et estimat for diameteren var 74.0362 og for spredningen var 0.00124.
Angiv et 95% konfidensinterval for μ.

Opgave 4.2

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

12.4	10.8	12.1	12.0	13.2	12.6	11.5	11.9	12.8	12.0

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

Opgave 4.3

Ved en batchproduktion er middeludbyttet 80 kg.
Laboratorieforsøg har antydet at vise ændringer i produktionsbetingelserne giver en forøgelse af udbyttet. For at undersøge det foretages 20 forsøg, hvis resultat viser, et gennemsnit på 83.2 kg med en spredning på1.5 kg.
a) Opstil et 95% konfidensinterval for middelværdien.
b) Viser det, at middeludbyttet sandsynligvis er steget?

Opgave 4.4

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

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

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

Opgave 4.5

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

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

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

Opgave 4.6

4 Konfidensinterval for normalfordelt variabel

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

pH	8.42	7.36	8.04	7.71	7.65	7.82

Den benyttede pH-målemetode antages på baggrund af tidligere lignende målinger at give normalfordelte resultater.
1) Angiv et estimat for opløsningens middelværdi og spredning.
2) Angiv et 95% konfidensinterval for pH.
3) Man finder, at radius i konfidensintervallet er for bredt.
   Angiv med tilnærmelse antallet af målinger der skal foretages, hvis radius skal være 0.1.

Opgave 4.7
De 10 øverste ark papir i en pakke med printerpapir har følgende vægt

4.21	4.33	4.26	4.27	4.19	4.30	4.24	4.24	4.28	4.24

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

Opgave 4.8
Til undersøgelse af alkoholprocenten i en persons blod foretages 4 uafhængige målinger, som gav følgende resultater (i ‰):

108	102	107	98

1) Opstil et 95% konfidensinterval for personens alkoholkoncentration.
2) Opstil et 95% konfidensinterval for målemetodens spredning.

5 HYPOTESETEST (ÉN NORMALFORDELT VARIABEL)

5.1 GRUNDLÆGGENDE BEGREBER

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

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

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

Eksempel 5.1. Hypotesetest med kendt spredning.
En fabrik har gennem mange år benyttet en metode, der på basis af en given mængde råmateriale gav et middeludbytte af et produceret stof på μ_0 = 69.2 kg og spredningen σ = 1.0 kg.
En nyansat ingeniør får til opgave at søge at forøge middeludbyttet ved en passende (billig) modifikation af procesbetingelserne.
Efter en række lovende eksperimenter i laboratoriet synes opgaven at være lykkedes, men det endelige bevis herfor er, ud fra et passende antal driftsforsøg statistisk at kunne "bevise", at middeludbyttet er blevet forøget.
Ud fra kendskab til de forskellige mulige støjfaktorer antages spredningen at være uændret på 1.0 kg.
Da driftsforsøgene er meget ressourcekrævende, bevilges der kun 12 delforsøg.
Der foretages 12 uafhængige delforsøg og udbyttet x måltes:

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

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

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

Løsning:

1) Løsningen opdeles for overskuelighedens skyld i en række trin

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

X = udbyttet ved den modificerede proces.

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

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

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

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

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

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

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

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

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

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

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

Hvis stikprøvens gennemsnit $\bar{x}$ er meget større end 69.2 kg (måske helt op mod 100 kg), så er der stor sandsynlighed for at udbyttet er steget. Man siger så, at **nulhypotesen forkastes**, eller at $\bar{x}$ ligger i forkastelsesområdet (se figur 5.1).

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

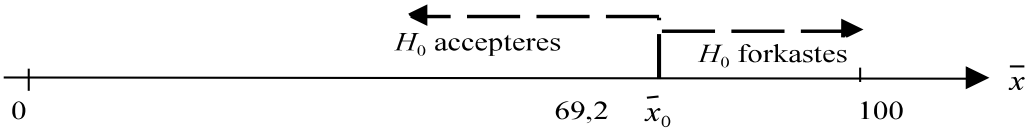


Fig. 5.1 *Accept- og forkastelsesområde*

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

Det er naturligvis ikke entydigt bestemt, hvad det vil sige, at noget er yderst usandsynligt. Man starter derfor enhver test med at fastlægge det såkaldte **signifikansniveau** α.

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

Da det kan have alvorlige økonomiske konsekvenser fejlagtigt at påstå at middeludbyttet er steget (produktionen omstilles osv.) ,så er man naturligvis interesseret i, at dette ikke sker.
Det normale i industriel produktion er, at sætte $\alpha = 5\%$, men er det eksempelvis medicinske forsøg, hvor det kan have alvorlige menneskelige konsekvenser, sættes α måske så lavt som 1% eller 0.1%, mens man i andre situationer måske sætter signifikansniveauet til 10%.
I dette eksempel er α sat til 5%.

1e) **Beregning af *P* - værdi**

Beregne gennemsnit:
Tallene indtastes i en søjle med navn x.
Analyze ▶ Distribution ▶ marker x og tryk på y,Columns ▶ ok
Der fremkommer et histogram og nogle tabeller.
Nederst ses en tabel hvor der står Mean
Man kan få alle de andre tal væk ved
rød pil ved Summary Statistics ▶ Customize Summery statistics ▶ Fjern de man ikke ønsker

$\bar{x} = 69.76$ kg.

Under forudsætning af at nulhypotesen H_0 : $\mu = 69.2$ kg er sand, så er $\overline{X}$ er normalfordelt med middelværdi $\mu_0 = 69.2$ og spredning $\frac{\sigma}{\sqrt{n}} = \frac{1.0}{\sqrt{12}} = 0.2887$.

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

=69.67

Da 69.76 > 69.67 ligger det målte gennemsnit altså i forkastelsesområdet.
Imidlertid vælger man i stedet at beregne den såkaldte P-værdi (Probability value) som er sandsynligheden for at få en værdi på det fundne stikprøvegennemsnit 69.76 eller derover, dvs. P-værdi = $P(\overline{X} \geq 69.76)$
Er denne P-værdi er mindre end α =0.05 må $\bar{x} = 69.76$ ligge i forkastelsesområdet (se figur 5.2)
Hvis P-værdien ligger over α ligger $\bar{x} = 69.76$ i acceptområdet, dvs. vi kan ikke bevise at middeludbyttet er steget.

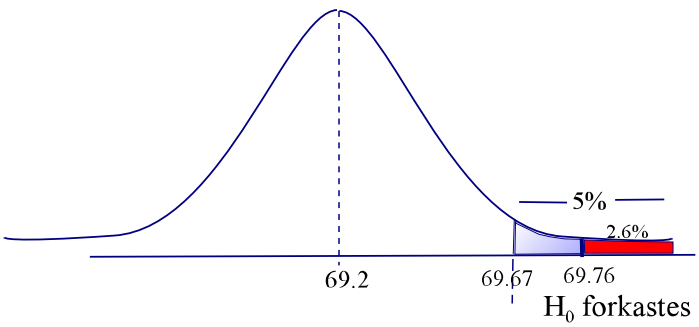


Fig 5.2 *P-værdi*

P værdi = $P(\overline{X} \geq 69.76) = 1 - P(\overline{X} < 69.76) =$ 1 - Normal Distribution(69,76, 69,2, 0,2887) = 0.0262

1f) **Konklusion**

Da P - værdi = 2.62% < 5% forkastes H_0 ,

Vi har et statistisk bevis for, at den modificerede proces giver et større middeludbytte.

Alternativt kunne vi have benyttet en testfunktion:
Tallene indtastes i en søjle med navn x.
Analyze ▶ Distribution ▶ marker x og tryk på y,Columns ▶ ok
Der fremkommer et histogram og nogle tabeller.
Rød pil ved x ▶ Test Mean ▶ udfyld menu med 69.2 og 1.
Da H: $\mu > 69.2$ skal vælges Prob>z, dvs. P-værdi=0.0265

	z Test
Test Statistic	1,9341
Prob > |z|	0,0531
Prob > z	0,0265*
Prob < z	0,9735

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

95% konfidensinterval:

Radius r = $z_{0.975} \cdot \frac{\sigma}{\sqrt{n}}$ = Normal Quantile(0,975) * $\frac{1}{\sqrt[2]{12}}$ =0.5658

95%Konfidensinterval [69.76-0.57 ; 69.76+0.57]=[69.19 ; 70.33]

At konfidensintervallet indeholder tallet 69.2 er klart i modstrid med at vi lige har vist, at middelværdien er større end 69.2.

Det skyldes, at konfidensintervallet forkaster med 2.5% til hver side, mens en ensidet test forkaster kun til en side med 5%.

Mere logisk ville det være, at lave en ensidet 95% konfidensinterval,

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

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

Alternativt kunne vi da de orginale tal findes i tabel have benyttet program i SAS.JMP:
Tallene indtastes i en søjle med navn x.
Analyze ▶ Distribution ▶ marker x og tryk på y,Columns ▶ ok
Der fremkommer et histogram og nogle tabeller.
Rød pil ved x ▶ Test Mean ▶ udfyld menu med 69.2 og 1.
Da H: $\mu > 69.2$ skal vælges Prob>z, dvs. P-værdi=0.0265

	z Test
Test Statistic	1,9341
Prob > |z|	0,0531
Prob > z	0,0265*
Prob < z	0,9735

95% konfidensinterval: lidt højere oppe ses en tabel, hvor man aflæser intervallet [69.24 ;70.28]

At tallene ikke er helt de samme skyldes at programmet ikke har så store afrundingsfejl.

Summary Statistics	
Mean	69,758333
Std Dev	0,8162646
Std Err Mean	0,2356353
Upper 95% Mean	70,276963
Lower 95% Mean	69,239704
N	12

Eksempel 5.2. Hypotesetest, hvor man får accept af H_0.

Samme problem som i eksempel 5.1, men nu er signifikansniveauet α =1%.

Løsning:

H_0: μ = 69.2 mod H: H_0: μ > 69.2

I eksemplet fandt vi på basis af 12 forsøg, at P-værdi = 2.6%.

Konklusion: H_0 accepteres , dvs.

vi kan **ikke** på et signifikansniveau på 1% bevise, at middelværdien var steget.

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

5.2 HYPOTESETEST MED UKENDT MIDDELVÆRDI OG SPREDNING

I eksempel 5.1 blev baggrunden for testen gennemgået. Samtidig antog vi, at spredningen var kendt eksakt. Dette er sjældent tilfældet, men havde vi haft over 30 målinger i stikprøven, ville det sædvanligvis være tilladeligt, at erstatte den eksakte værdi med den beregnede spredning s, og foretage de samme beregninger
Imidlertid er det på et statistisk program lige så let at lave den korrekte t-test, så det vil man nok altid gøre.

Eksempel 5.3 Ensidet hypotesetest om middelværdi (originale data ikke kendt)

Samme problem som i eksempel 5.1 men nu er spredningen ikke kendt eksakt.
På basis af 12 målinger beregnes gennemsnittet til 69.76 og spredningen til 0.82.
Kan man bevise på et signifikansniveau på 5% , at middeludbyttet er blevet forøget, dvs, er større end 69.2 kg?

Løsning

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

Som bekendt gælder $x_p = t_p \cdot \frac{s}{\sqrt{n}} + \mu \Leftrightarrow t_p = \frac{x_p - \mu}{\frac{s}{\sqrt{n}}}$ dvs. $t_p = \frac{69.76 - 69.2}{\frac{0.82}{\sqrt{12}}} = 2.3657$

Heraf følger (se evt. oversigt 5.4)

P-værdi = P(T>2.3672)= 1 - t Distribution(2,3657 , 11) = 0.0187

Da P-værdi < 5% forkastes H_0 , dvs. vi har et statistisk bevis for, at den modificerede proces giver et større middeludbytte.
95% konfidensinterval: Metoden kan ses i eksempel 5.5

Eksempel 5.4. Ensidet hypotesetest om middelværdi (originale data kendt)

Samme problem som i eksempel 5.1, men nu er spredningen ikke kendt eksakt.

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

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

Løsning:

1) X = udbyttet ved den modificerede proces.

X antages at være approksimativt normalfordelt $n(\mu, \sigma)$.
H_0 : μ = 69.2 kg. H: μ > 69.2 kg.
Tallene indtastes i en søjle med navn x.
Analyze ▶ Distribution ▶ marker x og tryk på y,Columns ▶ ok
Der fremkommer et histogram og nogle tabeller.
Rød pil ved x ▶ Test Mean ▶ udfyld menu med 69.2
Blandt en del tabeller findes

	t Test
Test Statistic	2,3695
Prob > |t|	0,0372*
Prob > t	0,0186*
Prob < t	0,9814

Da H: $\mu > 69.2$ skal vælges Prob>t, dvs. P-værdi=0.0186

Da P-værdi < 5% forkastes H_0, dvs. vi har et statistisk bevis for, at den modificerede proces giver et større middeludbytte.

2) I samme udskrift lidt højere oppe findes følgende tabel

Summary Statistics	
Mean	69,758333
Std Dev	0,8162646
Std Err Mean	0,2356353
Upper 95% Mean	70,276963
Lower 95% Mean	69,239704
N	12

Konfidensinterval [69.24 ; 70.28]

◆

Eksempel 5.5 Tosidet hypotesetest om middelværdi (spredning ikke kendt eksakt).
Ved fremstilling af et bestemt levnedsmiddel er det vigtigt, at et tilsætningsstof findes i levnedsmidler i en koncentration på 8.40 (g/l).
For at kontrollere om tilsætningsstoffet har en koncentration på ca. 8.40, udtager levnedsmiddelkontrollen 6 prøver af levnedsmidler. Resultaterne var:

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

Det ønskes på denne baggrund undersøgt om koncentrationen har den ønskede værdi.
Signifikansniveau sættes til 5%.
Løsning:
Lad X være koncentrationen af tilsætningsstoffet i levnedsmidlet.
Det antages, at X er normalfordelt $n(\mu, \sigma)$
Da det både er uønsket, at koncentrationen er for lille og at den er for stor, bliver nulhypotesen H_0: μ = 8.4 mod H: $\mu \neq 8.4$, dvs. vi har en tosidet test.
Bemærk, at selv om man vel egentlig hellere ville bevise, at koncentrationen er 8.4 og derfor helst ville have denne påstand i den alternative hypotese, er dette ikke muligt, da nulhypotesen skal indeholde et lighedstegn.
Løsning
Tallene indtastes i en søjle med navn x.
Analyze ▶ Distribution ▶ marker x og tryk på y,Columns ▶ ok
Der fremkommer et histogram og nogle tabeller.
Rød pil ved x ▶ Test Mean ▶ udfyld menu med 8.4
Der fremkommer bl.a. følgende tabel:

	t Test
Test Statistic	-1,3742
Prob > |t|	0,2118
Prob > t	0,8941
Prob < t	0,1059

Da vi har en tosidet test er P-værdi = 0.2118

Da P-værdi = 0.2118 > 0.5 accepteres nulhypotesen , dvs. vi kan ikke bevise, at koncentrationen afviger signifikant fra 8.4 g/l
Bemærk, at SAS.JMP har multipliceret P-værdi med 2, hvilket nok skyldes, at så skal vi altid sammenligne med 5%. ◆

Eksempel 5.6. Test af spredning

En virksomhed ønsker at kontrollere om den spredning som en bestemt målemetode angiver saltindholdet i en opløsning er under 0.25. 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

Med et signifikansniveau på 5% kan man da vise, at målemetoden ikke opfylder det stillede krav?

Løsning

Lad X = saltindhold i opløsning.

X antages normalfordelt $n(\mu,\sigma)$, hvor såvel μ som σ er ukendte.

H_o: $\sigma = 0.25$ imod H: $\sigma > 0.25$

Tallene indtastes i en søjle med navn x.

Analyze ▶ Distribution ▶ marker x og tryk på y,Columns ▶ ok

Der fremkommer et histogram og nogle tabeller.

Rød pil ved x ▶ Test Std Dev ▶ udfyld menu med 0.25

Der fremkommer bl.a. følgende tabel:

Test Standard Deviation	
Hypothesized Value	0,25
Actual Estimate	0,31623
DF	11
Test	**ChiSquare**
Test Statistic	17,6
Min PValue	0,1827
Prob < ChiSq	0,9087
Prob > ChiSq	0,0913

Da H: $\sigma > 0.25$ ses ved Prob>ChiSq dvs. P-værdi = 0.0913

Da P-værdi=9.13% > 5 %, accepteres H_0, dvs. det er ikke påvist, at spredningen ved målingerne er for stor, men der er dog nær ved at være signifikans. ♦

Eksempel 5.7. Test af spredning (uden original data)

En fabrikant af læskedrikke har købt en automatisk "påfyldningsmaskine".

Ved købet af maskinen har man betinget sig, at rumfanget af den påfyldte væske i middel skal have en spredning, der ikke overstiger 0.20 ml.

Efter kort tids anvendelse får man mistanke om, at spredningen er for stor. Mange klager over underfyldte flasker.

Derfor foretages en kontrol, hvor man tilfældigt udtager 20 flasker med læskedrik, og måler rumfanget af væsken i flasken. Det viser sig, at stikprøvens spredning er s = 0.24 ml.

Med et signifikansniveau på 5% er det da et statistisk bevis for, at den nye maskine ikke opfylder det stillede krav?

Løsning:

Lad X = rumfang af drik i flaske.

X antages normalfordelt $n(\mu,\sigma)$, hvor såvel μ som σ er ukendte.

H_o: $\sigma = 0.2$ imod H: $\sigma > 0.2$,

$\chi^2 = \frac{(n-1)\cdot s^2}{\sigma_0^2}$ (se oversigt 5.4) dvs. i det foreliggende tilfælde $\chi^2 = \frac{(20-1)\cdot 0{,}24^2}{0{,}2^2}$ =27.36

P- værdi = $P(\chi^2 \geq 27.36) =$ 1 - ChiSquare Distribution(27,36, 19) = 0.0965

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

♦

5.3. FEJL AF TYPE I OG TYPE II:

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

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

Det må være et krav til en god test, at der kun er en lille sandsynlighed for at begå en fejl af type I eller type II.
I eksempel 5.1 ville en type I fejl være, hvis man konkluderer, at den modificerede proces giver et større udbytte, selv om det ikke er tilfældet. Virksomheden bruger måske millionbeløb på at omlægge produktionen, og det er ganske forgæves.
En type II fejl ville være, at man ikke opdager, at den modificerede proces giver et større udbytte. Dette er naturligvis uheldigt, men hvis det skyldes, at forbedringen ikke blev opdaget, fordi den er ganske ringe, har det muligvis ingen praktisk betydning.
Hvis en test har signifikansniveau α og den beregnede P-værdi < α så **forkastes H_0 .**
Vi ved hermed, at P(type I fejl) $\leq \alpha$, dvs. vi rimelig sikre på, at have foretaget en korrekt beslutning. P-værdien angiver jo sandsynligheden for at vi træffer en forkert beslutning.
Hvis α = 5% og P-værdien er 4.25% forkastes H_0. Det samme sker, hvis P-værdi = 0.001%, men vi er her unægtelig noget sikrere på, at vi at vi træffer en korrekt beslutning.
Hvis vi **accepterer H_0** er det blot udtryk for, at vi ikke kan forkaste(svag konklusion: "H_0 frikendes på grund af bevisets stilling").
Man kan have begået en type II fejl, dvs. ikke opdaget, at den alternative hypotese var sand.

Eksempel 5.8. Fejl af type 2

Samme problem som i eksempel 5.1, men nu er signifikansniveauet α =1%

Løsning:

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

Som det ses af eksempel 5.8, så vil en formindskelse af muligheden for at begå en type 1 fejl (α formindskes) forøge sandsynligheden for at begå en type 2 fejl.
Den eneste måde hvorpå begge kan formindskes er at øge antallet n af forsøg.
Problemet hermed er, at man derved måske opdager en så lille forbedring, at det ikke er rentabelt at foretage en dyr ændring af fremstillingsprocessen.
Først når udbyttet overstiger en **bagatelgrænse** Δ vil man reagere.

Dimensionering af forsøg (vælge stikprøvestørrelse *n*).

Lad os antage, at virksomheden i eksempel 5.1 finder, at hvis stigningen i udbyttet ved den modificerede proces er mindre end Δ = 0.5 kg, så har det ingen praktisk interesse (Δ = 0.5 kg er bagatelgrænsen), og derfor gør det intet, hvis man ikke opdager det (begår en type II fejl).
Hvis derimod stigningen Δ er større end 0.5 kg, så har det stor betydning, og sandsynligheden for at begå en type II fejl må derfor være lille. Lad os sætte den til højst β = 10%.
Problemet er nu, hvor stor en stikprøvestørrelse n (antallet af delforsøg) der skal udføres, for at ovennævnte krav er opfyldt.
En sådan vurdering kaldes en **dimensionering** af forsøget. Udfører man det ud fra en dimensionering nødvendige antal forsøg, vil en accept af nulhypotesen nu betyde, at nok kan udbyttet være steget, men ikke så meget, at det har praktisk interesse.
I oversigt 5.4 er angivet de formler, der skal anvendes ved en dimensionering.
De følgende 2 eksempler viser anvendelsen heraf.

Eksempel 5.9. Dimensionering (kendt spredning).

Inden man i eksempel 5.1 begyndte at lave de dyre delforsøg, vil ingeniøren gerne have en vurdering af, hvor mange driftsforsøg der er nødvendige, når det vides, at det først er økonomisk rentabelt at gå over til den nye metode, hvis middeludbyttet er steget med mindst 0.5 kg.

1) Find stikprøvestørrelsen n, i det tilfælde, hvor Δ = 0.5 kg og β = 10%.
   Det antages stadig, at σ = 1.0 kg og signifikansniveauet er α = 5 %.

Lad n være den i spørgsmål 1 fundne stikprøvestørrelse.

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

Løsning

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

   Da testen er ensidet fremgår det af oversigt 5.4 at: $n \geq \left(\frac{z_{1-\alpha} + z_{1-\beta}}{\frac{\Delta}{\sigma}} \right)^2 = \left(\frac{z_{0.95} + z_{0.90}}{\frac{0.5}{1.0}} \right)^2$

$$= \left(\frac{\text{Normal Quantile}(0{,}95) + \text{Normal Quantile}(0{,}9)}{\frac{0{,}5}{1}} \right)^2 = 34.25 \qquad \underline{n = 35}$$

2a) H_0: μ = 69.2 mod H: H_0: μ > 69.2

$$P(\bar{x} > 69.8) = 1 - P(\bar{x} \leq 69.8) = 1 - \text{Normal Distribution}\left(69{,}8, 69{,}2, \frac{1}{\sqrt[2]{35}}\right) = 0.000193$$

Da P-værdi =0.000193<0.05 forkastes H_0: μ = 69.2 kg , dvs. vi er på et Signifikansniveau på 5% sikre på at middelværdien er over 69.2 kg.
Imidlertid kan vi ikke være sikre på at den er over bagatelgrænsen 69.2 + 0.5 = 69.7 kg
Lad H_0: μ = 69.7 mod H: H_0: μ > 69.7
Vi finder på samme måde som ovenfor, at P-værdi = 27.7%, dvs. en påstand om at middeludbyttet ligger over 69.7 kg vil være fejlagtig i ca. 28% af tilfældene.
Vi vil derfor næppe på den baggrund gå over til den nye metode.

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

Eksempel 5.10. Dimensionering, (ukendt spredning)

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

Hvor mange forsøg skal der mindst foretages, hvis α = 5%, β = 10%, Δ = 0.10 mg/l og et overslag over hvor stor σ er sætter denne til 0.15 mg/l.

Løsning:

Lad X = indhold af A (i mg/l) efter benyttelse af den ny metode.

X antages normalfordelt $n(\mu, \sigma)$, hvor såvel μ som σ er ukendte.

Da indholdet af stoffet A ønskes formindsket, bliver

nulhypotesen $H_0 : \mu = 1.7$ mg/l mod $H: \mu < 1.7$ mg/l, dvs. vi har en ensidet test.

Da σ ikke er kendt (kun et løst skøn kendes), er testen en t - test.

Formlen i oversigt 5.4 anvendes:

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

$\left(\frac{\text{Normal Quantile}(0,95) + \text{Normal Quantile}(0,9)}{\frac{0,1}{0,15}} \right)^2$ =19.27

Da n =20 < 30 bør man nu løse en ligning (se nedenfor)

Da spredningen jo var usikker, så vil man nok nøjes med at sætte n = 30

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

søjler benævnes n og n1. I søjle n skrives 20 og i søjle n1 skrives

$19,27 * \left(\frac{\text{t Quantile}(0,95, n-1)}{\text{Normal Quantile}(0,95)} \right)^2$

Derefter skrives 21 og 22

Resultat

n	n1
20	21,2953069...
21	21,1867096...
22	21,0891548...

Den ønskede dimensionering kræver altså 22 forsøg.

5.4. OVERSIGT over centrale formler i kapitel 5

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

Oversigt over test af middelværdi μ

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

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

Forudsætninger	Alternativ hypotese H	*P* - værdi	Beregning	H_0 forkastes
σ **ukendt.** $t = \frac{(\bar{x}-\mu_0)\cdot\sqrt{n}}{s}$.	$H: \mu > \mu_0$	$P(T \geq t)$	1-t Distribution(t,f) eller t-test	*P* - værdi $< \alpha$
	$H: \mu < \mu_0$	$P(T \leq t)$	t Distribution(t,f) eller t-test	
	$H: \mu \neq \mu_0$	$P(T \geq t)$ for $\bar{x} > \mu_0$ $P(T \leq t)$ for $\bar{x} \leq \mu_0$	som række 1 som række 2	*P* - værdi $< \frac{1}{2}\alpha$ dog hvis t-test : *P* - værdi $< \alpha$
σ **kendt eksakt**	$H: \mu > \mu_0$	$P(Y \geq \bar{x})$	1-normaldistribution $(\bar{x}, \mu_0, \sigma)$ se eksempel 5.1	*P* - værdi $< \alpha$
	$H: \mu < \mu_0$	$P(Y \leq \bar{x})$	normaldistribution $(\bar{x}, \mu_0, \sigma)$	
	$H: \mu \neq \mu_0$	$P(Y \geq \bar{x})$ for $\bar{x} > \mu_0$ $P(Y \leq \bar{x})$ for $\bar{x} \leq \mu_0$	son række 1 som række 2	*P* - værdi $< \frac{1}{2}\alpha$ dog hvis Z-test : *P* - værdi $< \alpha$

Dimensionering

$\Delta = |\mu - \mu_0|$ er den mindste ændring i μ der har praktisk interesse.

α =*P*(type I fejl), β = *P*(type II fejl)

Forudsætning	Hypotese	Formel	Beregning
σ **kendt eksakt**	Ensidet	$n \geq \left(\frac{z_{1-\alpha} + z_{1-\beta}}{\frac{\Delta}{\sigma}}\right)^2$	se eksempel 5.8
	Tosidet	$n \geq \left(\frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{\frac{\Delta}{\sigma}}\right)^2$	
σ er ukendt, men erstattes i formlerne af det bedste estimat eller gæt for spredningen.	Ensidet	$n \geq \left(\frac{z_{1-\alpha} + z_{1-\beta}}{\frac{\Delta}{\sigma}}\right)^2 \cdot \left(\frac{t_{1-\alpha}(n-1)}{z_{1-\alpha}}\right)^2$	Løse ligning, se eksempel 5.9

	Tosidet	$n \geq \left(\frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{\frac{\Delta}{\sigma}} \right)^2 \cdot \left(\frac{t_{1-\frac{\alpha}{2}}(n-1)}{z_{1-\frac{\alpha}{2}}} \right)^2$

Oversigt over test af varians σ^2

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

Forudsætning	Alternativ hypotese H	*P* - værdi	Beregning	H_0 forkastes
μ **ukendt** $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$	$H: \sigma^2 > \sigma_0^2$	$P(Q \geq \chi^2)$	se eksempel 5.6 eller 5.7	*P*-værdi< α
	$H: \sigma^2 < \sigma_0^2$	$P(Q \leq \chi^2)$		
	$H: \sigma^2 \neq \sigma_0^2$	$P(Q \geq \chi^2)$ for $\chi^2 \geq n-1$ $P(Q \leq \chi^2)$ for $\chi^2 < n-1$	som række 1 som række 2	*P*-værdi< $\frac{1}{2}\alpha$
μ **kendt** $\chi^2 = \frac{(n-1)s^2 + n(\bar{x}-\mu)^2}{\sigma_0^2}$	$H: \sigma^2 > \sigma_0^2$	$P(Q \geq \chi^2)$		*P*-værdi< α
	$H: \sigma^2 < \sigma_0^2$	$P(Q \leq \chi^2)$		
	$H: \sigma^2 \neq \sigma_0^2$	$P(Q \geq \chi^2)$ for $\chi^2 \geq n-1$ $P(Q \leq \chi^2)$ for $\chi^2 < n-1$	som række 1 som række 2	*P*-værdi< $\frac{1}{2}\alpha$

OPGAVER

Opgave 5.1

Et levnedsmiddel ("corned beef") forhandles i pakker på 100 g.
Ved fabrikationen tilsættes traditionelt et konserveringsmiddel B (nitrit).
Da man har mistanke om, at B anvendt i større mængder kan have uønskede bivirkninger, må der højst tilsættes 2.5 mg B pr. 100 g.
Fabrikanten reklamerer med, at der i middel højst er 2 mg B pr. pakke og spredningen er eksakt $\sigma = 0{,}30$ mg.
En konkurrent tvivler herpå, og vil teste påstanden.
Der købes i forskellige butikker i alt 36 pakker, og indholdet af B blev målt.
Man fandt et gennemsnit af B på $\bar{x} = 2.10$ mg.
Idet det antages at spredningen $\sigma = 0.3$, kan man ud fra disse data bevise på signifikansniveau $\alpha = 0.01$, at reklamen lyver.

Opgave 5.2

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

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

Opgave 5.3

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

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

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

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

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

Opgave 5.4

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

8.2	8.3	7.9	8.2	7.8	8.6	8.9	7.8	8.2

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

Opgave 5.5

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

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

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

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

Opgave 5.6

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

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

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

Opgave 5.7

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

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

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

Opgave 5.8

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

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

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

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

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

Opgave 5.9

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

1) Bestem den nødvendige stikprøvestørrelse *n,* for at man ved en forskel i fedtindhold på Δ = 0.30 gram højst har, at *P* (type I fejl) = α = 0.01 og *P* (type II fejl) = β = 0.05. ($\sigma \approx 0.42$ gram).
2) Man finder et gennemsnit på 13.1 gram og et estimat s for spredningen på 0.42 gram. Kan man ud fra disse data bevise på signifikansniveau α = 0.01, at middelindholdet afviger fra 14 gram? (Husk altid at anføre: Hvad *X* er. Antagelser. Nulhypotese. Beregninger. Konklusion.).
3) Er det på basis af resultaterne muligt at vurdere, om at en eventuel afvigelse er større end bagatelgrænsen på 0,4 gram.

6. REGNEREGLER FOR SANDSYNLIGHED, KOMBINATORIK

6.1 REGNEREGLER FOR SANDSYNLIGHEDER

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

Eksempel 6.1. Gennemgående eksempel.
To skytter Anders og Brian skyder hver ét skud mod en skydeskive. Sandsynligheden for at Anders rammer skiven er 0.80 mens Brian har en træfsandsynlighed på 0,60.
Et eksperiment består i at de hver skyder et skud.
Lad A være hændelsen at Anders rammer skiven og lad B være sandsynligheden for at Brian rammer skiven.
Vi har derfor, at P(A) = 0.80 og P(B) = 0.60. ♦

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

Fællesmængden til A og B benævnes $A \cap B$ og er mængden af alle udfald i udfaldsrummet U, der tilhører **både A og B** (Den skraverede mængde i figur 8.1).
Eksempelvis er $A \cap B$ i eksempel 6.1 hændelsen, **at både Anders og Brian** rammer skiven

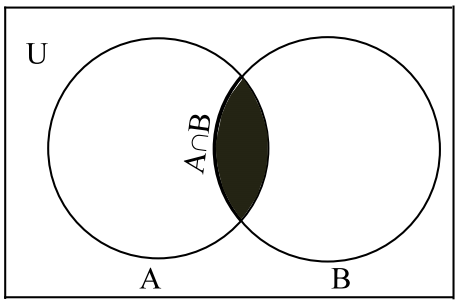


Fig 6.1. *Fællesmængde*

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 6.2)
Eksempelvis er $A \cup B$ i eksempel 8.1 den hændelse, at enten rammer Anders eller også rammer Brian skiven eventuelt gør de det begge.
Man kunne også udtrykke det ved at mindst en af dem rammer skiven.

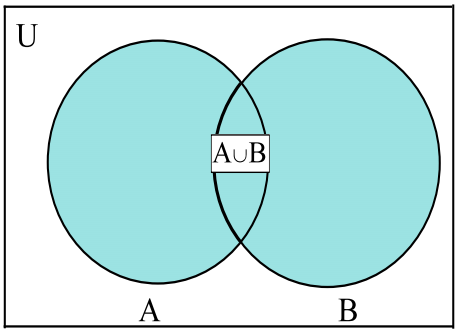


Fig. 6.2 *Foreningsmængde*

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

Sætningen fremgår umiddelbart ved at betragte arealerne i figur 6.3.

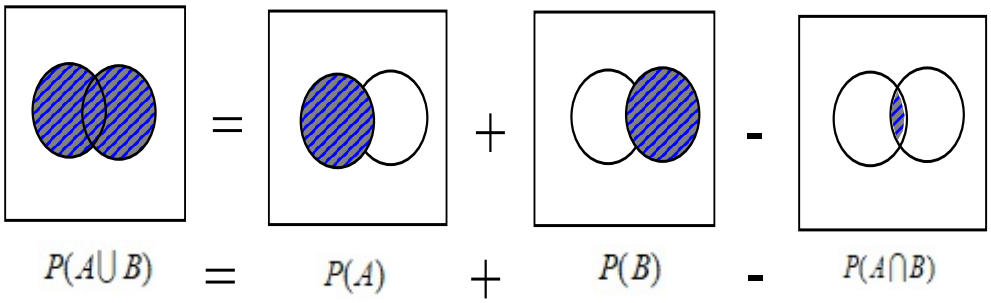


Fig.6.3 *Additionssætning*

Statistisk uafhængighed.
To hændelser A og B siges at være **statistisk uafhængige**, såfremt sandsynligheden for, at den ene hændelse indtræffer, ikke afhænger af, om den anden hændelse indtræffer.
I eksempel 6.1 må man eksempelvis antage, at om Anders rammer skiven har ingen indflydelse på om Brian rammer, så her må man antage A og B er uafhængige.

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

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

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

Eksempel 6.2 (eksempel 6.1 fortsat)
Lad A være hændelsen, at Anders rammer skiven, og lad B være hændelsen, at Brian rammer skiven. Det er givet, at P(A) = 0.80 og P(B) = 0.60.
Find sandsynligheden for
a) At både Anders og Brian rammer skiven
b) At enten Anders eller Brian (evt. begge) rammer skiven, dvs. mindst en af dem rammer skiven.
c) At hverken Anders eller Brian rammer skiven
Løsning:
a) Da hændelserne antages at være uafhængige gælder ifølge produktsætningen $P(A \cap B) = 0.8 \cdot 0.6 = \underline{\underline{0.48}}$
b) Ifølge additionssætningen gælder $P(A \cup B) = 0.6 + 0.8 - 0.48 = \underline{\underline{0.92}}$
c) $P(\overline{A} \cap \overline{B}) = P(\overline{A}) \cdot P(\overline{B}) = (1-0.8)(1-0.6) = \underline{\underline{0.08}}$ ♦

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

Er hændelserne A og B ikke uafhængige, kan man som beskrevet i afsnit 6.3 udlede en mere generel produktsætning

6.2. Betinget sandsynlighed

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

Eksempel 6.3. Ikke uafhængige hændelser
En fabrik har erfaring for, at den daglige produktion af glasfigurer indeholder 10 % misfarvede, 20% har ridser, og 1 % af produktionen er både ridsede og misfarvede.
Et eksperiment består i tilfældigt at udtage en glasfigur af produktionen. Lad A være hændelsen at få en misfarvet og lad B være hændelsen at få en ridset.
Her er $P(A) \cdot P(B) = 0.1 \cdot 0.2 = 0.02 \neq P(A \cap B) = 0.01$.

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

For at forklare den følgende definition, vil vi simplificere eksempel 6.3, idet vi antager, at den daglige produktion er 100 glasfigurer. I så fald er der 10 misfarvede figurer, 20 ridsede figurer, og 1 figur der er både misfarvet og ridset.
Hvis vi begrænser vort udfaldsrum til A, så er

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

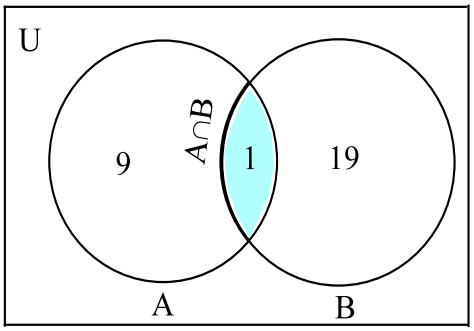


Fig. 6.4 *Taleksempel*

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

Benyttes produktsætningen på eksempel 6.1 fås $P(A \cap B) = P(A) \cdot P(B|A) = 0.1 \cdot 0.1 = 0.01$.

Eksempel 6.4: Betinget sandsynlighed.
En beholder indeholder 3 røde og 3 hvide kugler. Vi udtrækker successivt 2 kugler fra urnen.
Vi betragter følgende 2 hændelser:
A: Den først udtrukne kugle er rød. B: Den anden udtrukne kugle er rød.
Beregn $P(A \cap B)$ hvis
1) kugleudtrækningen foregår, ved at den først udtrukne kugle lægges tilbage før den anden udtrækkes.
2) kugleudtrækningen foregår, ved at den først udtrukne kugle **ikke** lægges tilbage før den anden udtrækkes.
Løsning
1) Her er $P(B|A) = \frac{3}{6}$ og derfor ifølge produktsætningen $P(A \cap B) = P(A) \cdot P(B|A) = \frac{1}{4}$
2) Her er $P(B|A) = \frac{2}{5}$ og derfor $P(A \cap B) = \frac{3}{6} \cdot \frac{2}{5} = \frac{1}{5}$

Bayes sætning

For to hændelser A og B for hvilken $P(A) > 0$ gælder "**Bayes sætning**" : $P(B|A) = \frac{P(B) \cdot P(A|B)}{P(A)}$

Bevis:

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

♦

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

Eksempel 6.5 (Bayes sætning)
I en officeruddannelse kan man vælge mellem en "teknisk" linie og en "operativ"linie. På en bestemt årgang har 60 % valgt den operative linie og af disse er 20% kvinder. På den tekniske linie er 10% kvinder.
Ved lodtrækning vælges en elev.
a) Find sandsynligheden for, at denne er en kvinde.
Ved ovenstående lodtrækning viste det sig at eleven var en kvinde.
b) Hvad er sandsynligheden for, at hun kommer fra den tekniske linie.
Løsning:
Vi definerer følgende hændelser:
T: Den udtrukne er tekniker
K: Den udtrukne er en kvinde.
a) $P(K) = P(T \cap K) + P(O \cap K) = P(K|T) \cdot P(T) + P(K|O) \cdot P(O) = 0.1 \cdot 0.4 + 0.2 \cdot 0.6 = \underline{0.16 = 16\%}$

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

En anden metode ville det være, at antage, at der bliver optaget 100 elever.
Vi har så følgende skema

	Kvinder	I alt
Operativ	12	60
Teknisk	4	40

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

6.3. Kombinatorik

6.3.1. Indledning:

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

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

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

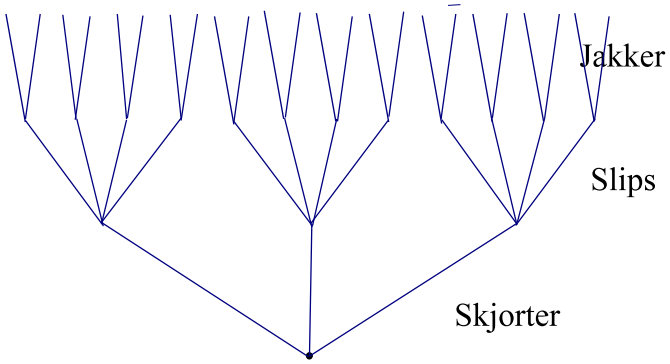
I sådanne tilfælde, bliver problemet derfor, hvorledes man let kan optælle antal udfald. Dette kan ofte gøres ved benyttelse af **kombinatorik**.

6.3.2. Multiplikationsprincippet

Multiplikationsprincippet: Lad et valg bestå af n delvalg, hvoraf det første valg har r_1 valgmuligheder, det næste valg har r_2 valgmuligheder, . . . og det n'te valg har r_n valgmuligheder.
Det samlede antal valgmuligheder er da $r_1 \cdot r_2 \cdot \ldots \cdot r_n$

Multiplikationsprincippet illustreres ved følgende eksempel.

Eksempel 6.6. Multiplikationsprincippet
En mand ejer 2 forskellige jakker, 4 slips og 3 forskellige fabrikater skjorter.
På hvor mange forskellige måder kan han sammensætte sin påklædning af skjorte,slips og jakke.
Løsning:
1) Valg af skjorte giver 3 valgmuligheder
2) Valg af slips giver 4 valgmuligheder
3) Valg af jakke giver 2 valgmuligheder
Ifølge multiplikationsprincippet giver det i alt $2 \cdot 3 \cdot 4 = \underline{\underline{24}}$ muligheder
Man kunne illustrere løsningen ved følgende "forgreningsgraf"



♦

Eksempel 6.7 Fakultet
På hvor mange måder kan 5 personer opstilles i en kø (i rækkefølge)
Løsning:
Pladserne i køen nummereres 1,2,3,4,5.
Plads nr. 1 i køen besættes 5 valgmuligheder
Plads nr. 2 i køen besættes 4 valgmuligheder
Plads nr. 3 i køen besættes 3 valgmuligheder
Plads nr. 4 i køen besættes 2 valgmuligheder
Plads nr. 5 i køen besættes 1 valgmulighed
I alt $5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$ forskellige rækkefølger.

Ved ***n* fakultet (*n* udråbstegn) forstås** $n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1$
Endvidere defineres 0! = 1.
SAS.JMP: Formula: **Trancendental ▶Factorial▶** 5 ▶ok

6.3.3 Ordnet stikprøveudtagelse

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

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

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

a) **Uden tilbagelægning**

Eksempel 6.8. Ordnet uden tilbagelægning

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

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

Løsning:

1) En formand vælges blandt 10 kandidater — 10 valgmuligheder
   En Kasserer vælges blandt de resterende 9 kandidater — 9 valgmuligheder
   Da der for hvert valg af formand er 9 muligheder for kasserer, følger af multiplikationsprincippet, at det totale antal forskellige bestyrelser er $10 \cdot 9 = \underline{\underline{90}}$.
2) Analogt fås ifølge multiplikationsprincippet at antal forskellige bestyrelser er
   $10 \cdot 9 \cdot 8 \cdot 7 = \underline{5040}$
   SAS.JMP : Formula ▶Trancendental ▶Factorial ▶indsæt 10 ▶division ▶Factorial ▶indsæt 6, ▶Ok
   Resultat $\frac{10!}{6!}$ =5040 ♦

Eksempel 6.8 begrunder følgende definition

Permutationer. Antal måder (rækkefølger eller "permutationer") som *m* elementer kan udtages (ordnet og uden tilbagelægning) ud af n elementer er

$$P(n,m) = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-m+1) = \frac{n!}{m!}$$

b) Med tilbagelægning

Eksempel 6.9. Ordnet, med tilbagelægning

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

Løsning:

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

6.3.4. Uordnet stikprøveudtagelse

Eksempel 6.10 Uordnet uden tilbagelægning

En beholder indeholdende 5 kugler med numrene k_1, k_2, k_3, k_4, k_5.
Vi udtager nu en stikprøve på 3 kugler uden tilbagelægning. Rækkefølgen kuglen tages op er uden betydning, dvs. der er ikke forskel på eksempelvis k_1, k_4, k_2 og k_1, k_4, k_2 .
Hvor mange forskellige stikprøver kan forekomme?

Løsning:

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

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

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

Antallet af kombinationer med *r* elementer betegnes C(*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 $C(n,r)=\frac{n!}{r!\cdot(n-r)!}$

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 $C(5,3)=\frac{5!}{3!\cdot 2!}$
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 C(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 $P(5,3)=C(5,3)\cdot 3!\Leftrightarrow C(5,3)=\frac{P(5,3)}{3!}\Leftrightarrow C(5,3)=\frac{5\cdot 4\cdot 3}{3!}=\frac{5\cdot 4\cdot 3\cdot 2\cdot 1}{3!\cdot 2!}=\frac{5!}{3!\cdot 2!}$ ◆

Eksempel 6.11. Antal kombinationer

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

Løsning:

Antal måder man kan sammensætte bestyrelsen er

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

Formula ▶Trancendental ▶ nChooseK ▶indsæt n = 10 , k=4 ▶Ok

Når man trykker på Ok skrives 210, og i formula står $\binom{10}{4}$ ◆

OPGAVER

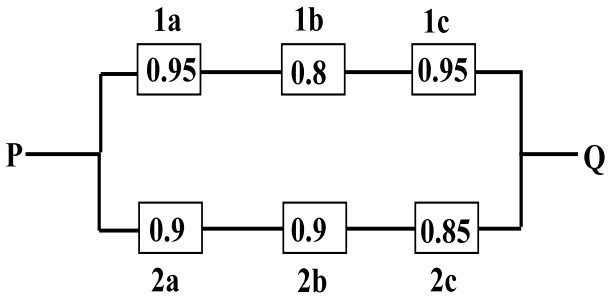
Opgave 6.1

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

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

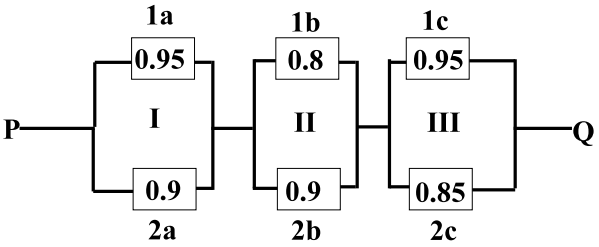
Opgave 6.2

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 6.3
Tre skytter skyder hver ét skud mod en skydeskive. De har træfsandsynligheder 0.75, 0.50 og 0.30.
Beregn sandsynligheden for
1) ingen træffere, 2) én træffer, 3) to træffere, 4) tre træffere.

Opgave 6.4
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 6.5
En klasse med 21 elever skal under en øvelse fordeles på 5 grupper. 4 af grupperne skal være på 4 elever, og 1 gruppe skal være på 5 elever.
På hvor mange måder kan fordelingen af eleverne på de 5 grupper foregå?

Opgave 6.6
Af en forsamling på 8 kvinder og 4 mænd skal udtages en arbejdsgruppe på 5 personer.
a) Gør rede for, at gruppen kan udvælges på 448 forskellige måder, når det forlanges, at den skal bestå af højst 3 kvinder og højst 3 mænd.
b) Beregn antallet af måder, hvorpå gruppen kan udvælges, når det forlanges, at de 5 personer ikke alle må være af samme køn.

Opgave 6.7
a) Bestem det antal måder, hvorpå bogstaverne A, B og C kan stilles rækkefølge.
b) Samme opgave for A, B, C, D, E ,F, G og H.

Opgave 6.8.
På et spisekort er opført 6 forretter, 10 hovedretter og 4 desserter.
1) Hvor mange forskellige middage bestående enten af forret og hovedret eller af hovedret og dessert kan man sammensætte.
2) Hvor mange forskellige middage bestående af en forret, en hovedret og en dessert kan man sammensætte.

Opgave 6.9
Bestem antallet af 5-cifrede tal, der kan skrives med to l-taller, et 2- tal og to 3-taller.

Opgave 6.10
Fire projektgrupper på en virksomhed antages at have sandsynlighederne 0.6, 0.7, 0.8 og 0.9 for at få succes med deres projekt. Grupperne antages at arbejde uafhængigt af hinanden. Find sandsynligheden for, at
a) alle grupper får succes,
b) ingen grupper får succes,
c) mindst 1 gruppe får succes,
d) i alt netop 1 gruppe får succes,
e) i alt netop 3 grupper får succes,
f) i alt netop 2 grupper får succes.

Opgave 6.11

En virksomhed fremstiller en bestemt slags apparater. Hvert apparat er sammensat af 5 komponenter. Heraf er 3 tilfældigt udvalgt blandt komponenter af typen a og 2 blandt komponenter af typen b. Det vides, at 10% af a-komponenterne er defekte og 20% af b-komponenterne er defekte. Et apparat fungerer hvis og kun hvis det ikke indeholder nogen defekt komponent.

Der udtages på tilfældig måde et apparat fra produktionen. Lad os betragte hændelserne:

A: Det udtagne apparat indeholder mindst 1 defekt a-komponent.

B: Det udtagne apparat indeholder mindst 1 defekt b-komponent.

1) Find $P(A)$, $P(B)$ og $P(A \cap B)$.
2) Find sandsynligheden for, at et apparat, der på tilfældig måde udtages af produktionen ikke fungerer.
3) Et apparat udtages på tilfældig måde fra produktionen og det konstateres ved afprøvning at det ikke fungerer. Find sandsynligheden for, at apparatet ikke indeholder nogen defekt a-komponent.

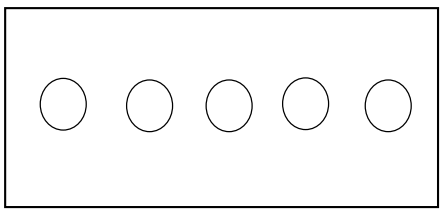
Opgave 6.12

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

Opgave 6.13

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

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



Opgave 6.14

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

Opgave 6.15

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

7. VIGTIGE DISKRETE FORDELINGER

7.1 INDLEDNING

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

7.2 HYPERGEOMETRISK FORDELING

Den "hypergeometriske fordeling", finder bl.a. anvendelse ved kvalitetskontrol af varepartier (jævnfør eksempel 7.3), ved markedsundersøgelser, hvor man uden tilbagelægning udtager en repræsentativ stikprøve på eksempelvis 500 personer

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

Eksempel 7.1. Hypergeometrisk fordeling
I en forening skal der blandt 5 kvindelige og 8 mandlige kandidater vælges en bestyrelse på 4 personer. Find sandsynligheden for, at der er netop 1 kvinde i bestyrelsen..
Løsning:
X = antal kvinder i bestyrelsen
At der skal være netop 1 kvinde i bestyrelsen forudsætter, at vi udtager 1 kvinde ud af de 5 kvinder og 3 mænd ud af de 8 mænd.
At udtage 1 kvinde ud af 5 kvinder kan gøres på C(5,1) måder
At udtage 3 mænd ud af 8 mænd kan gøres på C(8,3) måder.
Antal gunstige udfald er ifølge multiplikationsprincippet $C(5,1)\cdot C(8,3)$
Det totale antal udfald fås ved at udtage 4 personer ud af de 13 kandidater
Dette kan gøres på C(13,4) måder.

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

Formula ▶Discrete Probability▶Hypergeometric Probability, udfyld menu

Hypergeometric Probability(13, 5, 4, 1) =0.3916

Karakteristisk for en hypergeometrisk fordeling er, at elementerne i udfaldsrummet (kugler i en beholder) kan opdeles i **to** grupper.
En opdeling kunne som i eksempel 7.1 være kvinder og mænd eller som i kvalitetskontrol være i defekte varer og ikke-defekte varer.

Lad os antage, at vi har en beholder med N kugler, hvoraf de K er røde og resten har en anden farve.
Der udtrækkes en stikprøve på n kugler **uden tilbagelægning**.
Lad X være antallet af røde kugler blandt de n kugler.

X er hypergeometrisk fordelt med parametrene N, K, n (kort skrevet h(N,K,n))
P(X = x) er sandsynligheden for at netop x kugler er røde blandt de n udtrukne kugler.

X siges at være **hypergeometrisk** fordelt med parametrene N, K, n , hvor

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

SAS.JMP : Hypergeometric Probability (N, K, n, x)

Formlen udledes på samme måde som det skete i eksempel 7.1
Sætte x = 0, 1, 2, ... finder vi forskellige værdier af tæthedsfunktionen.

I "Supplement til statistiske grundbegreber" afsnit 7A bevises, at den hypergeometriske fordeling har middelværdien $E(X)=n\cdot p$ og spredningen $\sigma(X)=\sqrt{n\cdot p\cdot(1-p)\cdot\frac{N-n}{N-1}}$, hvor $p=\frac{K}{N}$.

Eksempel 7.2: Hypergeometrisk fordeling *h* (20, 9, 6).

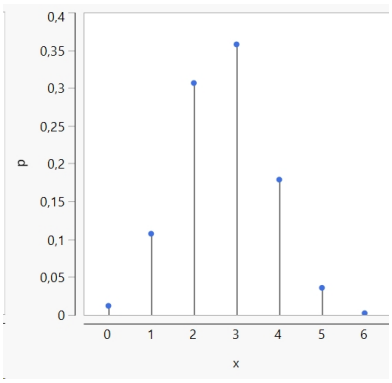
I en urne findes 20 kugler, hvoraf 9 er sorte, 11 er hvide.
Vi betragter det tilfældige eksperiment: "Udtrækning af en kugle og observation af farven på kuglen". Eksperimentet gentages 6 gange, idet den udtrukne kugle ikke lægges tilbage mellem hver udtrækning.
Lad X betegne antallet af udtrukne sorte kugler.
Find og skitser tæthedsfunktionen for X, og beregn middelværdi og spredning for X.
Løsning:
X er en diskret stokastisk variabel, der som er hypergeometrisk fordelt h (20, 9, 6) med tæthedsfunktionen $f(x) = P(X = x)$:
En søjle kaldes x og der indsættes 0

En søjle kaldes p og der indsættes Hypergeometric Probability(20, 9, 6, x)

derefter indsættes i søjle x tallet 1 osv.
Der fremkommer følgende tabel

x	p
0	0,0119195046
1	0,1072755418
2	0,306501548
3	0,3575851393
4	0,1787925697
5	0,0357585139
6	0,0021671827

Vælg Graph ▶Chart ▶Vælg p og tryk på Statistics ▶Data ▶tryk på x og indsæt i Catagories ▶ok ▶Rød pil ved Chart ▶Vertical Chart og Y Options ▶Needle Chart



Sættes $p = \frac{K}{N} = \frac{9}{20}$ er middelværdien $E(X) = n \cdot p = 6 \cdot \frac{9}{20} = \underline{\underline{2.7}}$ og

spredningen $\sigma(X) = \sqrt{n \cdot p \cdot (1-p) \cdot \frac{N-n}{N-1}} = \sqrt{6 \cdot \frac{9}{20} \cdot \left(1 - \frac{9}{20}\right) \cdot \frac{20-6}{20-1}} = \underline{1.046}$

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

Eksempel 7.3: Stikprøveudtagning (kvalitetskontrol)

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

Løsning:

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

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

SAS.JMP: Formula ▶ Discrete Probability p = Hypergeometric Distribution(600, 10, 25, 1) = <u>0.9389</u>

7.3 BINOMIALFORDELING

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

Eksempel 7.4. En binomialfordelt variabel.

En drejebænk producerer 1 % defekte emner.
Lad X være antallet af defekte blandt de næste 5 emner der produceres.
Vi ønsker at finde sandsynligheden for at finde netop 2 defekte blandt disse 5, det vil sige $P(X = 2)$.

Løsning:

Lad et eksperiment være at udtage et emne fra produktionen.
Resultatet af eksperimentet har to udfald: defekt, ikke defekt.
Eksperimentet gentages 5 gange uafhængigt af hinanden.
Der er en bestemt sandsynlighed for at få en defekt, nemlig $p = 0.01$.
Lad d være det udfald at få en defekt, og $\bar{d}$ være det udfald at få en fejlfri.

Vi opskriver nu samtlige forløb, der giver 2 defekte ud af 5 .

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

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

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

Det næste forløb må have sandsynligheden

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

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

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

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

Lad X være antallet af succeser blandt de n gentagelser
X er en diskret stokastisk variabel med tæthedsfunktionen

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

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

Eksempel 7.5 = eksempel 7.4 fortsat
1) X er binomialfordelt b(n,p) hvor n = 5 og p = 0.01
SAS.JMP: Formula ▶ Discrete Probability

P(X=2) = Binomial Probability (0,01 , 5 , 2) = 0.00097

Eksempel 7.6 Binomialfordeling

En drejebænk producerer 1 % defekte emner.
Lad X være antallet af defekte blandt de næste 100 emner der produceres.
Find sandsynligheden for at finde mindst 2 defekte blandt disse 100, det vil sige $P(X \geq 2)$.

Løsning

P(X ≥ 2)=1-P(X ≤ 1) 1 - Binomial Distribution(0,01, 100, 1) = = 0.2642 =26.42%

SÆTNING 7.1. (*middelværdi og spredning for binomialfordeling*).
Lad X være binomialfordelt b (*n, p*).
Der gælder da $E(X) = n \cdot p$ og $\sigma(X) = \sqrt{n \cdot p \cdot (1-p)}$.

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

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

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

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

Idet $X = X_1 + X_2 + \ldots + X_n$ er binomialfordelt *b* (*n*, *p*) fås af linearitetsreglen (kapitel 2 afsnit 2.4), at
$E(X) = E(X_1) + E(X_2) + E(X_3) + \ldots + E(X_n) = p + p + p + \ldots + p = n \cdot p$.
Endvidere fås af kvadratreglen i kapitel 1 afsnit 5, idet vi har uafhængige gentagelser, at
$V(X) = V(X_1) + V(X_2) + \ldots + V(X_n) = p \cdot (1-p) + p \cdot (1-p) + \ldots + p \cdot (1-p)$,
eller $V(X) = n \cdot p \cdot (1-p)$.

Eksempel 7.7: Tæthedsfunktion for binomialfordelt variabel .

Lad der på to af sidefladerne på en terning være skrevet tallet 1, på to andre sideflader være skrevet tallet 2 og på de sidste to sideflader være skrevet tallet 3.Vi betragter det tilfældige eksperiment:

"7 kast med en terningen og observation af det fremkomne tal.

Lad X betegne antallet af toere ved de 7 kast. X antages at være binomialfordelt $b\left(7, \frac{1}{3}\right)$.

1) Angiv tæthedsfunktionen $f(x)$ for X, og tegn et stolpediagram for $f(x)$.
2) Find middelværdi og spredning for X

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

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

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

Løsning:

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

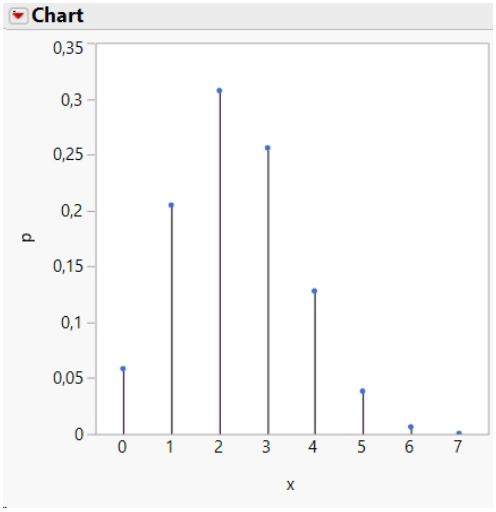
En søjle kaldes x og der indsættes 0

En søjle kaldes p og der indsættes Binomial Probability $\left(\frac{1}{3}, 7, x\right)$

derefter indsættes i søjle x tallet 1 osv.

x	p
0	0,0585276635
1	0,2048468221
2	0,3072702332
3	0,2560585277
4	0,1280292638
5	0,0384087791
6	0,0064014632
7	0,0004572474

Vælg Graph ▶Chart ▶Vælg p og tryk på Statistics ▶Data ▶tryk på x og indsæt i Catagories ▶ok ▶Rød pil ved Chart ▶Vertical Chart og Y Options ▶Needle Chart



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

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

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

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

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

Approksimation af hypergeometrisk fordeling med binomialfordeling.

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

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

Hypotesetest for binomialfordelt variabel.

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

Konfidensintervaller.

Ved løsning af en passende ligning (se oversigt 7.6) kan man finde de eksakte grænser for konfidensintervallerne.

De følgende to eksempler viser beregning af test og konfidensintervaller.

Eksempel 7.8. Ensidet binomialfordelingstest.

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

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

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

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

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

Løsning:

Lad X = antallet af rigtige svar.
X er binomialfordelt b (n, p), hvor n = 24 og p er ukendt.

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

P - værdi = $P(X \geq 13)$ = 1 - P(X ≤ 12)= 1 - Binomial Distribution(1/3, 24, 12) = 0.0284 = 2.84%

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

Eksempel 7.9. Konfidensinterval for parameteren *p* i binomialfordeling.

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

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

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

Løsning:

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

   Ifølge oversigt 7.8 er et estimat for p: $\tilde{p} = \frac{x}{n} = \frac{85}{100} = \underline{0.85}$

2) Approksimativ metode:

N-G = Binomial Quantile(0,85, 100, 0,025) / 100 = 0.78 Ø-G = Binomial Quantile(0,85, 100, 0,975) / 100 = 0,92

Resultat: 95% konfidensinterval : [0.78 ; 0.92]

Udenfor et 95% konfidensinterval ligger 5%, og af symmetrigrunde ligger der 2,5% på hver side.
Binomial Quantile giver som resultat det hele tal for hvilket P(X ≤ 0.025) (henholdsvis P(X ≤ 0,975), så forskellen mellem den korrekte værdi og den fundne kan højst være ca 1. Da vi derefter skal dividere med n = 100 bliver i dette tilfælde afvigelsen højst ca 0.01 som jo ikke afvige meget fra det korrekte.

Eksakt løsning:

Benyttes formel i oversigt 7.5.
Øvre grænse: Løs ligningen $P(X \leq 85) = 0.025$ med hensyn til p.
Benyttes "gættemetoden", dvs, kalde der en søjle p, en anden søjle "øvre"

og i formula under øvre skriver Binomial Distribution(p, 100, 85) * 100
fås ved at vælge p-værdier

0,915	2,1854510718
0,914	2,3982588709
0,913	2,6264478957
0,9135	2,5103911778
0,91355	2,4990026479

Heraf ses, at øvre grænse er 91.36%

Hvis man tilsvarende i en søjle "nedre" skriver fås

0,775	2,3345716831
0,776	2,4637849864
0,7765	2,5307102679
0,7763	2,5037515865
0,77627	2,4997295438

dvs. nedre grænse er 77.63%
95% Konfidensinterval: [77.63%; 91.36%]
Bemærk, at konfidensintervallet ikke ligger helt symmetrisk omkring 0.85, da binomialfordelingen ikke er helt symmetrisk omkring 0.85

Forklaring på formlen:
Udenfor et 95% konfidensinterval ligger 5%, og af symmetrigrunde ligger der 2,5% på hver side.
Er den sande værdi for p eksempelvis 90% vil der i middel være 90 ud af 100 overlevende beholdere. Nu fandt vi kun 85 ud af 100. Sandsynligheden P(X≤85) for at få 85 eller færre overlevende beholdere ud af 100 er derfor ret ringe. Vi har P(X≤85) = 0.0726. Selv om 7.26% er et lille tal, så er det dog over 2.5%, så en P-værdi på 90% ligger inde i konfidensintervallet.
For at finde den øvre grænse må vi derfor lade p stige indtil P(X≤85) = 0.025 .
Dernæst findes nedre grænse ved at lade *p* falde, indtil $P(X \geq 85) \approx 0.025$

Som det ses, er forskellen i forhold til den approksimative løsning meget lille, og som forventet ligger den eksakte løsning under den approksimative løsning.

7.4 POISSONFORDELINGEN

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

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

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

SÆTNING 7.2 (*Poissonfordeling*). *Lad X være en stokastisk variabel, som angiver antallet af impulser i et givet tidsrum (eller areal, volumen, produktionsenhed osv.), idet ethvert tidspunkt i tidsrummet har samme mulighed for at være impulstidspunkt som ethvert andet tidspunkt. Endvidere skal impulserne indtræffe tilfældigt og uafhængigt af hinanden* *).
Hvis det gennemsnitlige antal impulser i tidsrummet er $\mu > 0$ *, så siges X at være Poissonfordelt p* (μ) *med sandsynlighedsfordelingen (tæthedsfunktionen) f(x) = P(X = x) bestemt ved*

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

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

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

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

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

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

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

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

En bevisskitse for sætningen kan ses i "Supplement til statistiske grundbegreber" afsnit 7.C.

Eksempel 7.10: Antal revner pr. meter i et tyndt kobberkabel.

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

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

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

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

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

Løsning:

X = antal revner i 1 meter kabel.

X antages Poissonfordelt p (μ). (idet vi med tilnærmelse kan antage, at betingelserne i sætning 7.2 er opfyldt (impuls er her ridser).

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

1) $P(X=0) = \frac{1.23^0}{0!} \cdot e^{-1.23} = \underline{\underline{0.292}}$.

   SAS.JMP: Poisson Probability(1,23,0) = $\underline{\underline{0.292}}$

2) $P(X \geq 2) = 1 - P(X \leq 1) =$ 1 - Poisson Distribution (1,23 ,1) = $\underline{\underline{0.3482}}$

3) Y = antal revner i 2 meter kabel.

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

   Vi har derfor $P(X \leq 4)$ = Poisson Distribution(2,46, 4) = $\underline{\underline{0.8965}}$

4) Der er i alt 94 revner i 12 kabelstykker på hver 10 meter. Et estimat for μ er derfor

$$\tilde{\mu} = \frac{94}{12} = \underline{\underline{7.83}}.$$

◆

Hypotesetest og konfidensintervaller for Poissonfordelt variabel.

I kapitel 5 gennemgik vi ved en række eksempler de grundlæggende begreber for hypotesetestning og konfidensintervaller for én normalfordelt variabel.
Disse begreber kan uændret overføres til hypotesetestning og konfidensintervaller for Possonfordelt variabel.
Har man rådighed over et program med kumuleret Poissonfordeling kan testene gennemføres eksakt. (se oversigt 7.6)

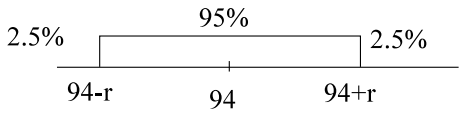
Eksempel 7.11. Ensidet Poissontest.

I eksempel 7.10 betragtede vi mikroskopiske revner i et kobberkabel. Fabrikken gik over til en anden og billigere produktionsmetode.

1) Test, om den nye metode giver færre revner end den gamle metode.
2) Forudsat, den nye metode giver signifikant færre revner end den gamle metode, skal man
   2a) Angiv et 95% konfidensinterval for middelværdien μ af antal revner pr. 120 meter kabel
   2b) Angiv et 95% konfidensinterval for middelværdien μ_1 af antal revner pr. 10 meter kabel.

Løsning:

1) Lad X betegne antallet af revner i 120 meter kabel ved ny metode
   X antages Poissonfordelt $p(\mu)$, hvor vi i eksempel 7.10 fandt at et estimat for μ var $\tilde{\mu} = 94$.
   Ved gammel metode er antal revner i 120 m kabel i middel $\mu_0 = 12.3 \cdot 12 = 147.6$
   Nulhypotese $H_0 : \mu = 147.6$ mod den alternative hypotese $H: \mu < 147.6$.
   P - værdi = $P(Y \le 94)$ = Poisson Distribution(147,6, 94) = $1.524 \cdot 10^{-6}$
   Da P - værdi < 0.05 forkastes nulhypotesen (stærkt) ,dvs. vi er sikre på, at middelantallet af revner er blevet formindsket ved at anvende den nye metode
   2a) oversigt 7.6 anvendes.
   120 m kabel: m = 94
   Øvre grænse x: Løs ligningen $P(X \le x) = 0.975$ med hensyn til x.
   Poisson Quantile(94,0.975) = 113
   Nedre grænse: Løs ligningen $P(X \le x) = 0.025$ med hensyn til x
   Poisson Quantile(94,0.025) = 75
   95% Konfidensinterval: [75.0; 113.0]
   Forklaring på formlen:
   Udenfor et 95% konfidensinterval ligger 5%, og af symmetrigrunde ligger der 2,5% på hver side. (jævnfør figuren)



   Jo mindre den sande værdi af middelværdien er, jo mindre er sandsynligheden for, at gennemsnittet blev 94.
   Vi leder derfor i grænsen efter et x , så $P(X \le x) = 0.025$

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

7.5 APPROKSIMATIONER

Undertiden kan man under visse forudsætninger med en rimelig nøjagtighed approksimere (erstatte) f.eks. en binomialfordeling eller en Poissonfordeling med en normalfordeling.
Dette kan give nogle simplere beregninger.

7.6. OVERSIGT over centrale formler i kapitel 7

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

Konfidensinterval

Forudsætninger	Estimat for p	100 (1 - α) % konfidensinterval for parameter
approksimation metode 1 max. fejl ca $\frac{100}{n}$ %	$\tilde{p} = \frac{x}{n}$	$\frac{BinomialQuantile(\tilde{p},n,0.025)}{n} \le p \le \frac{BinomialQuantile(\tilde{p},n,0.975)}{n}$ se eksempel 7.9
eksakt	$\tilde{p} = \frac{x}{n}$	nedre grænse: Løs ligning $P(X \ge x) = \frac{\alpha}{2}$ med hensyn til p. Gættemetode: se eksempel 7.9 øvre grænse: Løs ligning $P(X \le x) = \frac{\alpha}{2}$ med hensyn til p

Test af parameter p for binomialfordelt variabel

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

Alternativ hypotese H	*P* - værdi	Beregning	H_0 forkastes
$H: p > p_0$	$P(Y \ge x)$	1-Binomial Distribution(p_0,n, x-1) se eksempel 7.8	*P*-værdi $< \alpha$
$H: p < p_0$	$P(Y \le x)$	Binomial Distribution(p_0,n, x)	
$H: p = p_0$	$P(Y \ge x)$ for $x > n \cdot p_0$ $P(Y \le x)$ for $x \le n \cdot p_0$	som række 1 som række 2	*P*-værdi $< \frac{1}{2}\alpha$

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

Der optælles i alt *m* impulser i en stikprøve

Konfidensinterval

Forudsætning	Estimat for parameter	100 (1 - α) % konfidensinterval for parameter
	μ = m	nedre grænse x x = Poisson Quantile(m, $\frac{\alpha}{2}$) øvre grænse x: x = Poisson Quantile(m, 1 - $\frac{\alpha}{2}$) se eksempel 7.1
Approksimation m ≥ 10	μ = m	$m - z_{1-\frac{\alpha}{2}}\sqrt{m} \le \mu \le m + z_{1-\frac{\alpha}{2}}\sqrt{m}$

Test af parameter μ for Poissonfordelt variabel .

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

Alternativ hypotese H	*P* - værdi	Beregning	H_0 forkastes
$H: \mu > \mu_0$	$P(Y \ge m)$	1 - Poisson Distribution(μ_0, m)	*P* - værdi $< \alpha$
$H: \mu < \mu_0$	$P(Y \le m)$	Poisson Distribution(μ_0,m) se evt. eksempel 7.1	
$H: \mu \ne \mu_0$	$P(Y \ge m)$ for $x \ge \mu_0$ $P(Y \le m)$ for $x < \mu_0$	som række 1 som række 2	*P*-værdi $< \frac{1}{2}\alpha$

OPGAVER

Opgave 7.1

Ved en lodtrækning fordeles 3 gevinster blandt 25 lodsedler. En spiller har købt 5 lodsedler.

1) Beregn sandsynligheden for at spilleren vinder netop én gevinst.

Lad den stokastiske variable X være bestemt ved

X = antal gevinster som spilleren vinder

2) Find og skitser tæthedsfunktionen for X

3) Beregn middelværdien for X

Opgave 7.2

Fra et sædvanligt spil kort udtrækkes på tilfældig måde 3 kort uden tilbagelægning. Bestem sandsynlighederne for hver af hændelserne

A: Der udtrækkes kun 8'ere.

B: Der udtrækkes lutter hjerter.

C: Der udtrækkes 2 sorte og 1 rødt kort.

Opgave 7.3

På en undervisningsinstitution skal 105 studerende holde fest sammen med deres 23 lærere. Et festudvalg på 5 personer vælges tilfældigt.

a) Beregn sandsynligheden for at der kommer netop 1 lærer med i udvalget.

b) Beregn sandsynligheden for at der kommer mere end 1 lærer med i udvalget

Opgave 7.4.

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

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

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

Opgave 7.5

En fabrikant fremstiller en bestemt type radiokomponenter. Disse leveres i æsker med 30 komponenter i hver æske. En køber har den aftale med fabrikanten, at hvis en æske indeholder 4 defekte komponenter eller derover, kan køberen returnere æsken, i modsat fald skal den godkendes. Køberen kontrollere hver æske ved en stikprøve, idet han af æsken udtager 10 komponenter tilfældigt. Lad X være antal defekte i stikprøven. Der overvejes nu to planer:

1) Hvis $X = 0$, så godkendes æsken, ellers undersøges æsken nærmere.

2) Hvis $X \leq 1$, så godkendes æsken, ellers undersøges æsken nærmere.

Hvad er sandsynligheden for, at en æske, der indeholder netop 4 defekte komponenter, bliver godkendt af køberen ved metode 1 og ved metode 2.

Opgave 7.6

En tipskupon har 13 kampe med 3 mulige tegn - 1, x og 2 - for hver kamp. En person bestemmer tegnet, der skal sættes for hver kamp, ved tilfældig udtrækning af en seddel fra 3 sedler med tegnene henholdsvis 1, x og 2.

Angiv sandsynligheden for, at personen opnår netop 8 rigtige tippede kampe på sin kupon.

Opgave 7.7

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

Opgave 7.8
En "sypigetipper" (M/K) deltog i tipning 42 gange i løbet af et år. På hver tipskupon var der 13 kampe, ved hver af hvilke tipperen ved systematisk gætning satte et af de 3 tegn: 1, x, 2. Beregn sandsynligheden p for, at tipperen det pågældende år tippede mindst 200 kampe rigtigt.

Opgave 7.9
Blandt familier med 3 børn udvælges 50 familier tilfældigt. Angiv sandsynligheden for, at der i mindst 8 af disse familier udelukkede er børn af samme køn.

Opgave 7.10.
Ved en fabrikation af plastikposer leveres disse i æsker med 100 poser i hver. Ved en godkendelseskontrol af et parti plastikposer udtages og undersøges en tilfældigt udtaget æske, og partiet godkendes, såfremt æsken højst indeholder én defekt pose.
Vi antager, at den løbende produktion af poser er således, at hver produktion med sandsynligheden 2% giver en pose, der er defekt; vi vil senere formulere dette således, at produktionen er **i statistisk kontrol** med fejlsandsynligheden $p = 2\%$.
Hvor stor er sandsynligheden for, at partiet under disse omstændigheder accepteres?

Opgave 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
Ved et køb af 10000 plastikbægre aftaltes med leverandøren, at det skal være en forudsætning for købet, at partiet godkendes ved en stikprøvekontrol.
Kontrollen udøves ved, at 200 bægre udtages tilfældigt af partiet og kontrolleres. Partiet godkendes, såfremt højst 5 af de 200 bægre er defekte.
Beregn sandsynligheden for, at partiet godkendes, hvis det i alt indeholder 300 defekte bægre.

Opgave 7.13
En fabrikant får halvfabrikata hjem i partier på 200000 enheder. Fra hvert parti udtages en stikprøve på 100 enheder og antallet af fejlagtige blandt disse noteres.
Hvis dette antal er mindre end eller lig med 2, accepteres hele partiet; i modsat fald undersøges partiet yderligere.
1) Hvad er sandsynligheden for, at et parti med en fejlprocent på 1 vil blive yderligere undersøgt.
2) Hvor stor er sandsynligheden for, at et parti med en fejlprocent på 5 vil blive accepteret.

Opgave 7.14
En maskinfabrikant påtænker at købe 100000 møtrikker af en bestemt type. Man beslutter sig til at købe et tilbudt parti af den nævnte størrelse, såfremt en stikprøve på 150 møtrikker højst indeholder 4% defekte møtrikker.
1) Beregn sandsynligheden for, at partiet bliver godkendt af maskinfabrikken, såfremt det indeholder
   a) 4% defekte møtrikker,
   b) 7,5% defekte møtrikker,
2) Bestem, for hvilken procentdel defekte møtrikker det ovennævnte parti har 50% sandsynlighed for at blive godkendt af maskinfabrikken.

Opgave 7.15
En ny vaccine formodes med en sandsynlighed på mindst 85% at have en forebyggende virkning over for en bestemt influenzatype.
Før en truende influenzaepedemi vaccineres et hospitalspersonale på 600 personer med den pågældende vaccine. 125 af disse bliver smittet af sygdommen.
Kan dette opfattes som en eksperimentel påvisning af, at vaccinen er mindre virksom end ventet?

Opgave 7.16
1) Antag, at en vis type af fostermisdannelse normalt forekommer med hyppigheden 164 tilfælde p. 100000 fødsler. Beregn sandsynligheden for 3 eller flere fostermisdannelser blandt 256 fødsler.
2) For at undersøge om forholdene i et bestemt arbejdsmiljø forøger hyppigheden af denne type misdannelse, undersøgte man hyppigheden af misdannelser for mødre, som under graviditeten havde haft den aktuelle type af arbejde, og fandt 3 misdannelser blandt 256 fødsler. Kan den forøgede relative hyppighed i dette materiale skyldes tilfældigheder?

Opgave 7.17
Udsættes planterne af en bestemt sort roser for meldugssmitte, bliver i middel brøkdelen p angrebet, hvor p er mindst 0.20. En rosengartner fremavler en rosenstamme, som han påstår er mere modstandsdygtig over for meldugssmitte. For at kontrollere denne påstand bliver 100 roser af den nye stamme udsat for meldugssmitte. Det viser sig, at 12 roser bliver angrebet.
1) Bekræfter dette resultat rosengartnerens påstand? (Husk altid at anføre: Hvad X er. Antagelser. Nulhypotese. Beregninger. Konklusion.).

Hvis rosengartneren har ret, skal man
2) Angiv et estimat $\tilde{p}$ for den nye stammes p.
3) Angiv et 95% konfidensinterval for den nye stammes p.

Opgave 7.18
En fabrikant af chip til computere reklamerer med, at højst 2% af en bestemt type chip, som fabrikken sender ud på markedet er defekte.
Et stort computerfirma vil købe et meget stort parti af disse chip, hvis påstanden er rigtigt. For at teste påstanden købes 1000 af dem. Det viser sig, at 33 ud af de 1000 er defekte.
Kan fabrikantens påstand på denne baggrund forkastes på signifikansniveau 5% ?

Opgave 7.19
En producent af billigt plastiklegetøj får mange klager over at en bestemt type legetøj er defekt ved salget. Legetøjet sælges til butikkerne i kasser på 10 stk, og som et led i en kvalitetetskontrol udtages 100 kasser og antallet x af defekt legetøj optaltes. Følgende resultater fandtes:

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

Lad p være sandsynligheden for at få et defekt stykke legetøj.
1) Find et estimat $\tilde{p}$ for p.
2) Angiv et 95% konfidensinterval for p.

Opgave 7.20
Af 1000 tilfældigt udvalgte patienter, der led af lungekræft, var 823 døde senest 5 år efter sygdommen blev opdaget.
Angiv på dette grundlag et 95% konfidensinterval for sandsynligheden for at dø af denne sygdom senest 5 år efter at sygdommen bliver opdaget.

Opgave 7.21
En fabrikant af lommeregnere vurderer, at ca. 1% af de producerede lommeregnere er defekte.
For at få en nøjere vurdering heraf ønskes udtaget en stikprøve, der er så stor, at radius i et 95% konfidensinterval for fejlprocenten p er højst 0.5%.
Find stikprøvens størrelse n.

Opgave 7.22
På en fabrik fremstilles gulvtæpper, som har størrelsen 20 m^2. Ved fabrikationen er der gennemsnitlig 6 vævefejl pr. 100 m^2 klæde.
1) Beregn sandsynligheden for, at et tilfældigt gulvtæppe ingen vævefejl har.
2) Beregn sandsynligheden for, at et tilfældigt gulvtæppe højst har 2 vævefejl.

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

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

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

Opgave 7.23
Et radioaktivt præparat undergår gennemsnitligt 100 desintegrationer (sønderdelinger) p. minut. Lad X betegne antal desintegrationer i et sekund (som er lille i forhold til præparatets halveringstid).
Find $P(X \le 1)$.

Opgave 7.24
Ved en TV-fabrikation optælles som led i en godkendelseskontrol antal loddefejl p. 5 TV-apparater. Fabrikanten ønsker at få et overblik over antal loddefejl, og optalte derfor antal loddefejl på 24 tilfældigt udtagne TV apparater. Resultatet fremgår af skemaet:

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

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

Opgave 7.25
På et teknisk universitet er et centralt edb-anlæg i konstant brug. Man har erfaring for, at anlægget i løbet af en 20 ugers periode har gennemsnitligt 7 maskinstop.
Beregn sandsynligheden p for, at anlægget i en 4 ugers periode har mindst ét maskinstop.

Opgave 7.26
På en fabrik indtræffer i gennemsnit 72 ulykker om året. Antag, at de forskellige ulykker indtræffer uafhængigt af hinanden, og at de er nogenlunde jævnt fordelt over året.
Beregn, idet et arbejdsår sættes lig med 48 uger, sandsynligheden for at der i en uge indtræffer flere end 3 ulykker.

Opgave 7.27
Til et bestemt telefonnummer er der i løbet af aftenen i middel 300 opkald i timen.
Beregn sandsynligheden for, at der i løbet af et minut er højst 8 opkald.

Opgave 7.28
En fabrikation af fortinnede plader finder sted ved en kontinuerlig elektrolytisk proces. Umiddelbart efter produktionen kontrolleres for pladefejl. Man har erfaring for, at der i middel er 1 pladefejl hvert 5'te minut.
Beregn sandsynligheden for, at der højst er 5 pladefejl ved en halv times produktion.

Opgave 7.29
Lastbiler med affald ankommer tilfældigt og indbyrdes uafhængigt til en losseplads. Lossepladsens maksimale kapacitet er beregnet til, at der i middel ankommer 90 lastbiler p. time. Ledelsen af pladsen føler, at travlheden er blevet større i den sidste tid, således at antallet af lastbiler overskrider den maksimale kapacitet. For at undersøge dette, foretages en optælling af lastbiler i perioder à 10 minutter. Følgende resultater fremkom:

14	17	18	16	18	12	22	16	21	18

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

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

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

3) Angiv et 95% konfidensinterval for middelværdien μ [lastbiler/time].

Opgave 7.30
Nedenstående tabel viser fordelingen af 400 volumenenheder med hensyn til antal gærceller pr. volumenenhed.

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

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

Opgave 7.31
Ved inspektion af en produktion med isolering af kobberledning taltes der i løbet af 50 minutter i alt 11 isoleringsfejl.
Idet antallet af isoleringsfejl p. 50 minutter antages at være Poissonfordelt $p(\mu_1)$, 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 p. 5 minutter antages at være Poissonfordelt $p(\mu_2)$, skal man
2a) angive et estimat for μ_2.
2b) angive et 95% konfidensinterval for μ_2.

8 ANDRE FORDELINGER

8.1 INDLEDNING

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

8.2 Kontinuerte fordelinger

8.2.1 DEN REKTANGULÆRE FORDELING

DEFINITION af rektangulær fordeling med parametrene a og b.

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

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

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

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

SÆTNING 8.1. (*Middelværdi og spredning for rektangulær fordeling*).

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

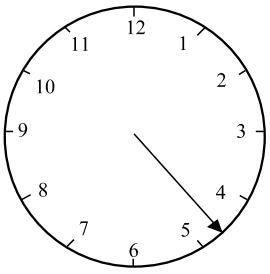
Bevis:

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

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

Eksempel 8.1 Kontinuert variabel.

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



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

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

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

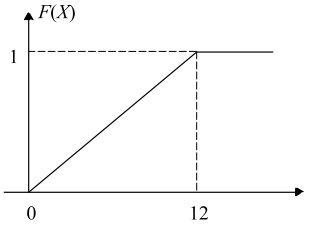
Angiv tæthedsfunktion og fordelingsfunktion for *X* og skitser disse.

Løsning:

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

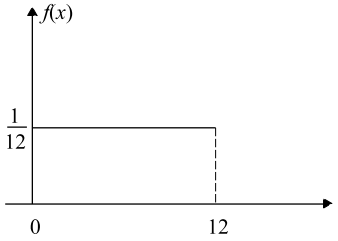
er fordelingsfunktionen for *X*

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



Ved differentiation fås tæthedsfunktionen

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



8.2.2 EKSPONENTIALFORDELINGEN

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

SÆTNING 8.2 .*Eksponentialfordeling*.

Lad W være en Poissonfordelt stokastisk variabel.

Lad det gennemsnitlige antal impulser i en tidsenhed være λ. Lad X være tiden indtil næste impuls.
X er da en kontinuert stokastisk variabel med sandsynlighedsfordelingen (tæthedsfunktionen)
$f(x) = P(X = x)$ bestemt ved

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

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

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

Bevis:

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

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

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

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

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

Bevis for middelværdi og spredning:

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

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

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

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

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

ε

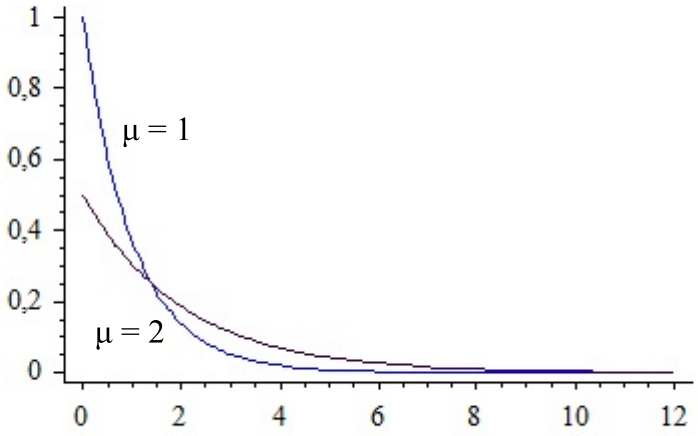


Fig 9.1 *Eksponentialfordelingerne* exp(1) *og* exp(2)

Eksempel 8.2. Afstanden mellem successive revner i kabel.

Vi betragter det i eksempel 7.11 omtalte problem, hvor man fandt, at antallet N af mikroskopiske revner i et kobberkabel er Poissonfordelt. Der var i gennemsnit 12.3 af den type revner pr. 10 meter. Lad X være afstanden mellem to på hinanden følgende revner.

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

Løsning

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

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

◆

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

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

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

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

◆

Eksempel 8.3. Levetid for elektriske pærer.

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

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

Løsning

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

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

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

◆

8.2.3 WEIBULLFORDELINGEN

Hvis komponenterne i et elektronisk apparat ikke "slides", dvs. den fremtidige levetid ikke afhænger af den foregående tid, er som nævnt i afsnit 9.3 eksponentialfordelingen velegnet som model for apparatets levetid.
Hvis derimod de pågældende komponenters eventuelle svigten afhænger af den forløbne tid, kan man ofte med fordel benytte den i det følgende nævnte **Weibullfordeling** som approksimativ model for apparatets levetid (model for apparatets pålidelighed).

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

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

siges at være <u>*Weibullfordelingen*</u> $wei(k, \mu)$.

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

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

Det ses, at Weibullfordelingen kan opfattes som en generalisation af eksponentialfordelingen, idet $wei(1, \mu) = \exp(\mu)$.
Såfremt levetiderne for komponenter i et apparat aftager jo længere tid apparatet har været i funktion (på grund af slid), kan man benytte en Weibullfordeling med $k > 1$ som approksimativ model for apparatets levetid.

8.2.4 DEN LOGARITMISKE NORMALFORDELING

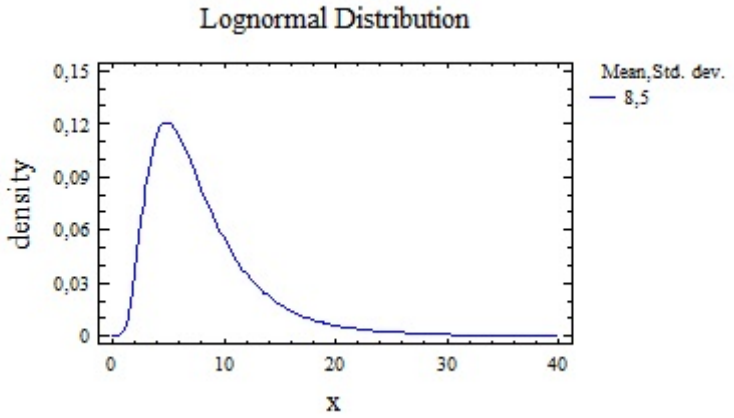
Indenfor det biokemiske eller biologiske område (forsøgsdyrs reaktionstid, cellevækst m.v.) er den stokastiske variabel X ikke normalfordelt, men hvis man foretager en logaritmisk transformation $Y = \ln X$ er *Y* (approksimativt) normalfordelt.
Man siger så, at *X* er **logaritmisk normalfordelt.**

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

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

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

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



8.3 Diskrete fordelinger

8.3.1 Den generaliserede hypergeometriske fordeling.

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

Eksempel 8.4. Generaliseret hypergeometrisk fordeling.

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

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

LØSNING:

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

Analogt med begrundelsen for den hypergeometriske fordeling fås:

$$P(X_1=2, X_2=3, X_3=1)=\frac{C(5,2)\cdot C(4,3)\cdot C(3,1)}{C(12,6)}=\frac{10\cdot 4\cdot 3}{924}=\underline{\underline{013}}$$

◆

8.3.2 Polynomialfordelingen.

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

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

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

DEFINITION af polynomialfordeling.

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

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

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

Dette illustreres ved følgende eksempel:

Eksempel 8.5. Polynomialfordelingen

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

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

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

LØSNING:

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

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

OPGAVER

Opgave 8.1

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

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

Det antages, at antallet N af telefonopkald til nummeret er Poissonfordelt. Lad T være tiden mellem to opkald.

1) Angiv fordelingsfunktionen for T, og giv et estimat for middelværdien μ.
   Vink: Antage, at for alle observationer i et interval er tidsrummet mellem observationerne intervallets midterværdi.
2) På baggrund af den i spørgsmål 1 fundne estimat for μ, ønskes bestemt $P(2 < T \leq 3)$.
3) Af tabellen ses, at i intervallet]2; 3] forekommer i alt 16 observationer. Angiv hvor mange observationer man må forvente, ud fra resultatet i spørgsmål 2.

Opgave 8.2

Om en bestemt type elektriske komponenter vides, at deres levetider er eksponentialfordelte med en middellevetid på 800 timer.

1) Find sandsynligheden for, at en komponent holder mindst 200 timer.
2) Find sandsynligheden for, at en komponent holder mellem 600 og 800 timer.
3) En komponent har holdt i 900 timer. Find sandsynligheden for, at den kan holde i mindst 200 timer mere.
4) I et elektrisk system indgår netop én komponent af denne type. Hver gang komponenten svigter, udskiftes den øjeblikkeligt med en ny komponent af samme type. Find sandsynligheden for, at komponenten udskiftes 12 gange i løbet af 8000 timer.

Opgave 8.3

Antag, at levetiderne for en bestemt slags elektroniske komponenter er uafhængige og alle er eksponentialfordelt med en middellevetid på 3 (år). Betragt et delsystem bestående af 3 sådanne komponenter i seriekobling:(en seriekobling ophører at fungere, når én af komponenterne ophører at fungere). Bestem middellevetiden for et sådant system.

Opgave 8.4

Nedbrydningstiden i den menneskelige organisme for et givet kvantum af et bestemt stof antages at være eksponentialfordelt med middelværdien 5 timer.
Ved et forsøg indsprøjtes stoffet samtidig i 10 patienter.

1) Beregn sandsynligheden (afrundet til et helt antal procent) for, at stoffet hos en tilfældig valgt patient vil være nedbrudt efter 8 timers forløb.
2) Beregn sandsynligheden for, at stoffet efter 8 timers forløb vil være nedbrudt hos mindst 5 af patienterne.
3) Efter hvor mange timers forløb vil der være ca. 90% sandsynlighed for, at stoffet er nedbrudt hos samtlige 10 patienter?
4) Hvor mange patienter skal indgå i en ny undersøgelse, hvis der skal være ca. 95% sandsynlighed for, at der er mindst en patient, hvis organisme efter 8 timers forløb endnu ikke har nedbrudt stoffet?

Opgave 8.5

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

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

Opgave 8.6

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

Opgave 8.7

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

9 STATISTISK PROCESKONTROL

9.1. INDLEDNING

Formålet med det meget omfattende begreb kvalitetsstyring (også kaldet kvalitetskontrol[1]) er at kontrollere, styre og forbedre kvaliteten af et produkt. Endvidere at nedsætte virksomhedens samlede styrings- og fejlomkostninger. Forskellige undersøgelser antyder, at mange virksomheder ved overgang til kvalitetsstyring kan nedsætte deres samlede kvalitetsomkostninger fra størrelsesordenen 10-15% af omsætningen til 3-5% af denne.

Der er på international basis opstillet krav, som skal opfyldes for opnåelse af en " kvalitetscertificering" af virksomheders kvalitetsstyring, bl.a. anført i serien af "ISO 9000-standarder". Virksomheder, som opfylder de pågældende krav, kan opnå et såkaldt "ISO 9000 - certifikat".

Vi begrænser os i nærværende kapitel til omtale af **statistisk proceskontrol** (SPC = Statistical Process Control) i egentligt materialeproducerende virksomheder.
Formålet med denne er at styre produktionsprocesserne således, at fejlproduktion forebygges.

En industriel produktionsproces kan formelt betragtes som en talfrembringende proces.
Eksempler herpå er angivet i det følgende skema:

nr	Produktion af	Frembragte tal
1	Kunstgødning	Kvælstofindhold
2	Jernbjælker	Trykstyrke , Siliciumindhold
3	TV-apparater	Antal loddefejl pr. apparat
4	Patroner	Antal defekte patroner i en stikprøve
5	Vin	Alkoholprocent
6	Flasker	Rumindhold
7	Film	0: defekt, 1: ikke defekt
8	Tekstil	Antal fejl pr. m^2
9	Aksler	Diameter

[1] På engelsk kaldet "Quality Control" eller "Quality Management" .

9.2 STATISTISK KONTROL.

Proces i statistisk kontrol. Ved enhver proces vil de frembragte tal udvise en "naturlig" variation, uanset hvor godt processen er planlagt og hvor omhyggeligt den bringes til udførelse og vedligeholdes. Denne "naturlige" variation der i praksis ikke kan kontrolleres, er et resultat af talrige små påvirkninger/variationsårsager ("common causes").
Hvis en proces **kun** er påvirket af tilfældige variationsårsager, siges den at være i **statistisk kontrol**.

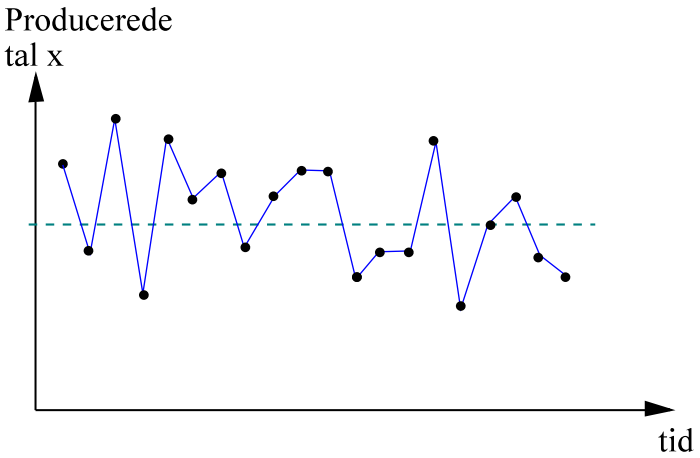


Fig 9.1. *Proces i statistisk kontrol*

De producerede tal er uafhængige observationer fra en population med en bestemt sandsynlighedsfordeling. Den dertil svarende statistiske variabel kaldes procesvariablen. Er en proces i statistisk kontrol kan man derfor på basis af en stikprøve estimere dens parametre, og kan eksempelvis beregne størrelsen af en kommende fejlproduktion.

Proces ude af statistisk kontrol. Udover den tilfældige variation kan der forekomme en variation, som kan tilskrives bestemte årsager, f.eks. maskiner, operatører eller råvarer; maskiner kan være indstillet forkert, en operatør kan være træt eller uopmærksom, et råvareparti kan være af dårlig kvalitet. Man taler i disse situationer om væsentlige, specielle eller påviselige variationsårsager ("assignable causes")
Hvis en proces i løbet af et givet tidsrum går fra at være i statistisk kontrol i en procestilstand til at være i statistisk kontrol i en anden procestilstand, har processen været **ude af statistisk kontrol** indenfor det pågældende tidsrum.

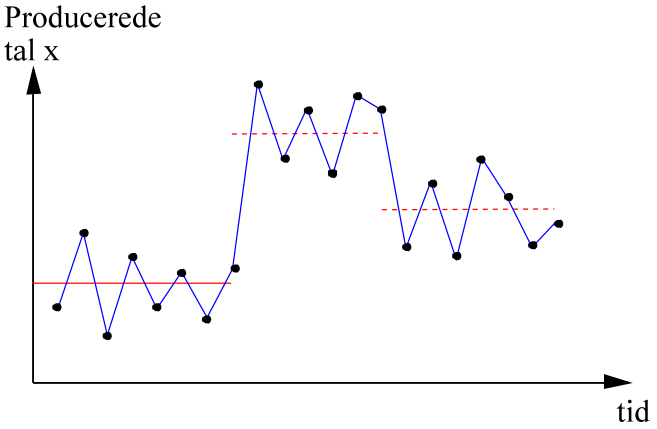


Fig.9.2. *Proces ude af statistisk kontrol*

9.3 OPBYGNING AF KONTROLKORT .

Kontrolkort (også kaldet Shewhart kontrolkort efter opfinderen) er baseret på, at man af den løbende produktion med regelmæssige mellemrum udtager stikprøver af størrelsen *n*.
Figur 9.3 viser en normalfordeling sammen med et enkelt kontrolkort.

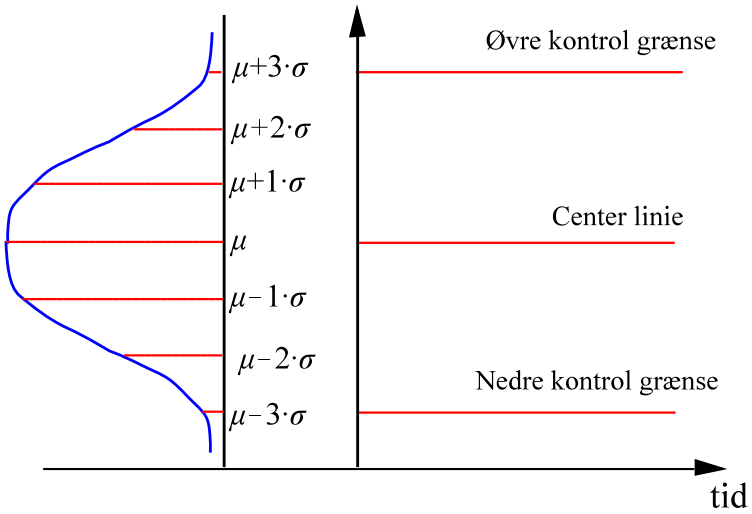


Fig 9.3. *Kontrolkort med øvre og nedre kontrolgrænser*

Hvis en på basis af stikprøven beregnet parameter eksempelvis gennemsnittet $\bar{x}$ holder sig indenfor nogle kontrolgrænser, antager man at processen er i kontrol med hensyn til denne parameter. Hvis den beregnede størrelse ligger udenfor kontrolgrænserne vil man "slå alarm" fordi så kan det tænkes at processen er kommet ud af kontrol.

Stikprøvernes størrelse *n*.
Generelt gælder det, at det er bedre at tage små stikprøver på 4-5 emner ud med korte mellemrum end at tage store stikprøver på 20-25 emner ud med lange mellemrum. Formålet er jo, at man ønsker hurtigt at opdage, hvis processen er ved at komme ud af kontrol (fordi eksempelvis en maskine er ved at gå i stykker). Bruger man store stikprøver vil man kunne opdage selv forholdsvis små forskydninger i niveauet, men på grund af at der går temmelig lang tid mellem man tager stikprøverne, kan en stor katastrofal forskydning blive opdaget for sent.

9.4. KONTROLKORTANALYSE.

Anvendelsen af kontrolkort kræver, at processen fra starten er i statistisk kontrol. Indførelsen af proceskontrol i en "ny" produktion kræver derfor, at man ved en såkaldt kontrolkortanalyse får undersøgt og om nødvendigt justeret processen således, at den kommer i statistisk kontrol. Først når dette er tilfældet, kan man estimere de ukendte parametrene og konstruere kontrolkortet.

Ved en kontrolkortanalyse foretages en række målinger af procesvariablen. Samtidig med en måling registreres en række oplysninger, f.eks. observationstidspunkt, temperatur, råmaterialeparti, arbejdshold m.v., samt hvilke personer der har foretaget målingerne. På grundlag heraf inddeles måltallene i "rationelle undergrupper", inden for hvilke de formodes at være produceret under samme væsentlige betingelser. Endvidere tilstræber man, at antal målinger er det samme i hver undergruppe.

Indsamling og opdeling af et observationsmateriale i rationelle undergrupper kræver ofte en betydelig teknisk indsigt i den betragtede produktionsproces. Alle de i praksis uundgåelige variationsårsager skal bidrage til den faktiske variation indenfor undergrupper, hvorimod alle de variationsårsager, som man mener kan være væsentlige variationsårsager, ikke må bidrage til variationen indenfor undergrupper.

De variationsårsager, som måske er væsentlige, må altså kun indvirke fra undergruppe til undergruppe. Mener man således, at tal produceret af forskellige maskiner kan være en væsentlig variationsårsag, så må disse tal ikke være placeret i samme undergruppe.
På basis af undergrupperne konstrueres så kontrolkort som beskrevet i eksempel 9.1. Falder en undergruppe udenfor de beregnede kontrolgrænser, så må man undersøge nærmere hvad årsagen kan være. Er det eksempelvis en gruppe hvis tal er produceret af en bestemt maskine, så må man ofre en hovedreparation på maskinen eller kassere den. På den måde får man ikke alene processen i kontrol, men man øger også kvaliteten af den igangværende proces.

Når kontrolkortet er lavet anvendes den jo i "den løbende kontrol", dvs. man foretager målinger eksempelvis hveranden time, og håber, at man så hurtigt opdager hvis processen er ved at komme ud af kontrol. I afsnit 9.6 er dette nærmere beskrevet.

Der udarbejdes forskellige typer kontrolkort afhængig af sandsynlighedsfordelingen for procesvariablen X.
I afsnit 9.5 behandles tilfældet hvor X approksimativt er en kontinuert normalfordelt variabel.
I afsnit 9.8 ser på de tilfælde, hvor X approksimativt er binomialfordelt eller Poissonfordelt.

Det er i alle tilfælde væsentligt, at observationerne er uafhængige, mens mindre afvigelser fra den forventede sandsynlighedsfordeling ikke vil give nogen væsentlig forøgelse i falske alarmer.

9.5 PROCESVARIABLEN ER NORMALFORDELT

9.5.1 Stikprøvestørrelse større end 1

Lad os antage at processen er i kontrol med en middelværdi på μ og en spredning på σ. Vi kender ikke de eksakte værdier, men ønsker at beregne estimater herfor.

For at kunne beregne kontrolgrænserne for $\bar{x}$ -kortet: $\bar{x} \pm 3\frac{\tilde{\sigma}}{\sqrt{n}}$ må man kende et estimat for spredningen σ. Man starter derfor altid med at lave et kontrolkort for spredningen. Man udtager i produktionen løbende ("af folk på gulvet") stikprøver. Da det tidligere var besværligt at beregne estimatet s , foretrak man ofte, at benytte variationsbredderne R_i som et mål for spredningen. Derfor vil man stadig møde mange " R -kontrolkort" selv om "s - kontrolkort" er blevet mere almindelige.

Sædvanligvis vil man benytte et statistikprogram som SAS.JMP til beregning af kontrolgrænser m.m., men i notatet "Supplement til Anvendt statistik" som findes på min hjemmeside er vist de tabeller og formler hvormed man kunne beregne dem ved "håndkraft".

Eksempel 9.1. Kontrol af stof i levnedsmiddelprodukt.

En levnedsmidddelvirksomhed har problemer med at holde koncentrationen af et bestemt stof A i et konservesprodukt mellem nogle tolerancegrænser. Man vælger derfor at få foretaget en kontrolkortanalyse. På basis af tidligere erfaringer inddeles målingerne i 30 undergrupper , som hver har deres karakteristika:(råvarecharge, apparatur, tidspunkt på dagen osv.).
Hver undergruppe er på 5 målinger.

Gruppe	Målinger					Gruppe	Målinger				
1	13	8	2	5	8	16	16	11	14	8	17
2	0	6	1	9	15	17	9	4	4	8	9
3	4	2	4	3	4	18	6	1	1	3	13
4	3	15	8	3	5	19	7	0	5	7	2
5	5	10	5	4	0	20	10	0	10	12	7
6	9	5	13	7	7	21	3	7	5	10	12
7	0	4	4	3	9	22	3	0	10	5	4
8	9	3	0	6	0	23	3	3	0	6	9
9	14	0	0	5	3	24	0	2	3	6	7
10	3	9	5	0	2	25	2	3	5	4	10
11	5	8	0	7	8	26	3	1	4	2	4
12	3	2	2	7	4	27	2	4	5	13	4
13	5	11	14	8	3	28	0	22	7	2	11
14	13	5	5	12	7	29	3	5	9	8	6
15	7	0	1	0	6	30	9	7	10	13	0

1) Beregn uden anvendelse af hjælpemidler $\bar{x}$ og R -værdi for gruppe 1
2) Foretag ved hjælp af $\bar{x}$ − og R - kort en kontrolkortanalyse og opstil kontrolkort, der kan benyttes til en løbende kontrol af indholdet af det undersøgte stof.

Løsning .

1) $\bar{x} = \frac{13+8+2+5+8}{5} = \frac{36}{5} = 7.2$ R= 13 - 2 = 11

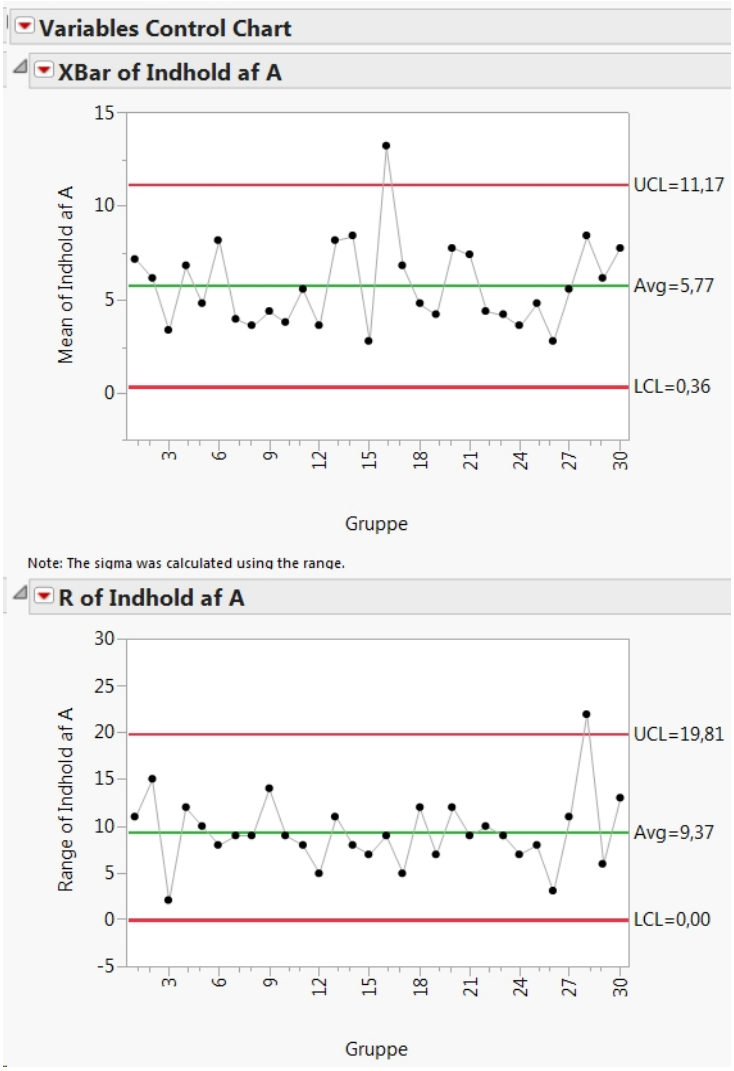
2) Data indtastes på sædvanlig måde:

Gruppe	Indhold af A
1	13
1	8
1	2
1	5
1	8
2	0
2	6
2	1
2	9
2	15
3	4
3	2
3	4

osv.

Vælg Analyze►Quality and Proces►Control Chart►X Bar►I menu vælg Process = Indhold af A ►Sample Label = gruppe►Marker "Xbar, R, kSigma►OK

Vi får følgende udskrift



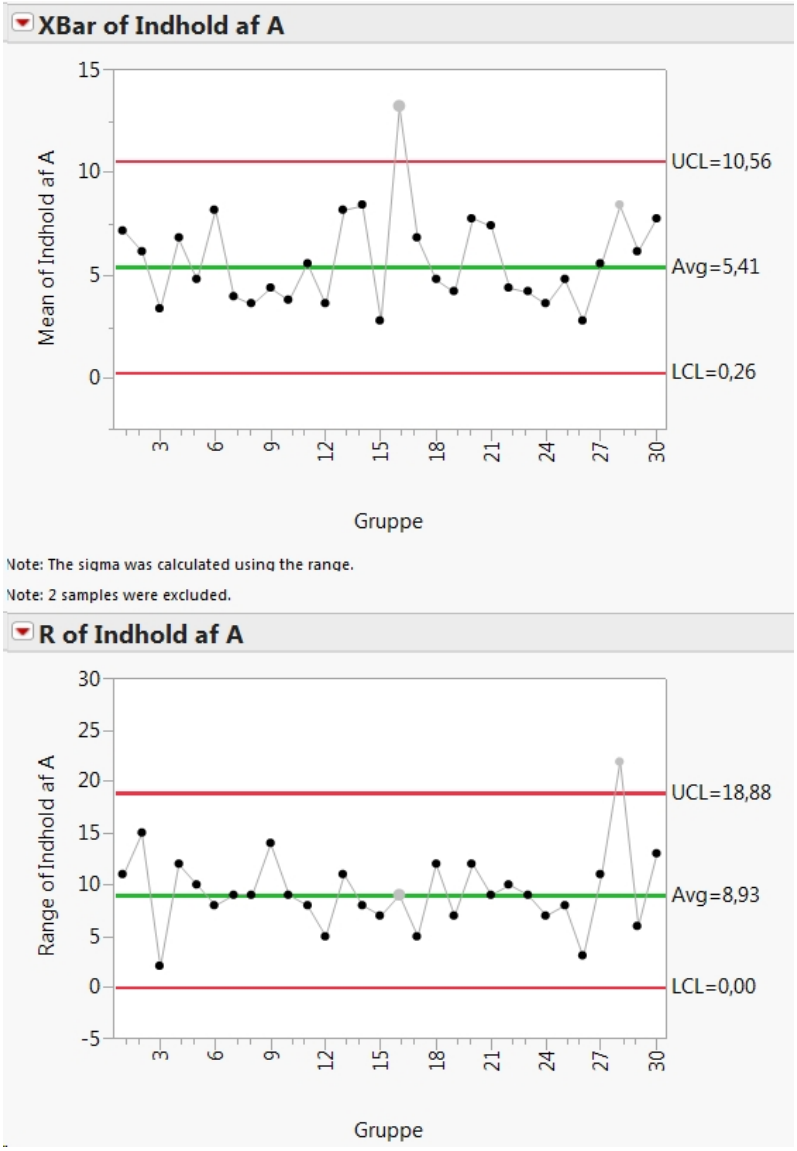
Først foretages en *R* - kort analyse.
Det ses, at gruppe 28 er udenfor kontrolgrænserne på R-kortet.
Man foretager nu en nærmere undersøgelse af hvilke forhold i gruppe 28 der kan tænkes at bevirke dette. Hvis det eksempelvis skyldes et bestemt apparat, kan man kassere dette eller reparere det.
Vi vil i det følgende udskyde gruppen, revidere grænserne og se om de resterende grupper nu falder indenfor de nye grænser.
Det kan somme tider være svært umiddelbart at se om et punkt falder indenfor eller udenfor kontrolgrænsen
Det er ikke tilfældet her, men ellers kan man gøre følgende :Rød pil ved R-kort ► Test beyond limits
Nu bliver alle punkter udenfor markeret
Synes man figuren er for lille og uoverskuelig, så
Højre musetast på figur ►Size/Scale►Y-axis►Angiv Min , Max og Incrediment.

133	27	5
134	27	13
135	27	4
136	28	0
137	28	22
138	28	7
139	28	2
140	28	11
141	29	3
142	29	5

Gruppe 28 udskydes.
Cursor placeres på gruppe 28 på R-kort ▶venstre musetast,
I datatabel markeres nu gruppe 28 med blåt ▶Cursor på gruppen ▶- Højre musetast ▶Exclude
Rød pil ved "Variable Control Chart ▶Redo ▶Redo analysis

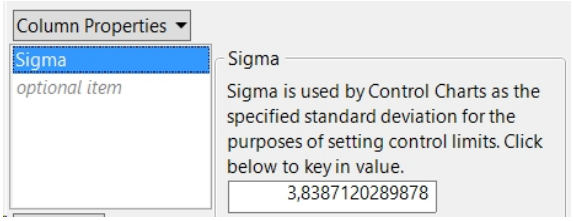
Vi får nye kontrolkort med nye grænser
For R-kortet er UCL = 18.88.
Vi ser, at nu er der ingen udenfor R-kortet, men stadig en gruppe (gruppe 16) udenfor kontrolgrænserne på $\bar{x}$ - kortet.
Vi udskyder nu dette punkt efter samme metode som før

Derefter er der ingen punkter på hverken $\bar{x}$ - kortet eller R- kortet, der er udenfor grænserne
De to kort kan nu benyttes til den løbende proceskontrol.



Kontrolgrænserne for middelværdien er [0.26 ; 10.56]

Ønskes et estimat for spredningen for en enkelt måling vælges
Vælg Rød pil ved "Variable Control Chart"▶Save Sigma
Placer cursor på overskriften i data "Indhold af A"▶højre musetast , Column Info



Et estimat for spredningen er 3.8387

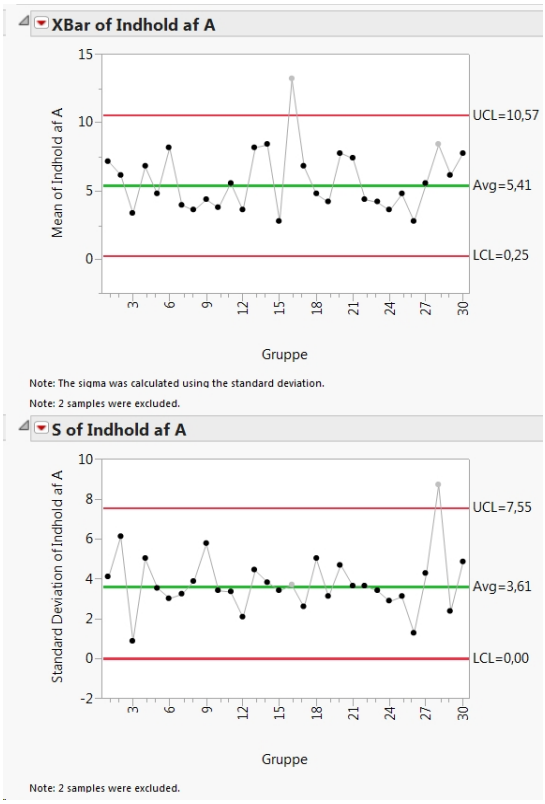
Eksempel 9.2. Kontrol af stof i levnedsmiddelproduktion.

Samme spørgsmål som i eksempel 9.1, men udarbejd i stedet for R - kortet et s- kort.

Løsning:

Vælg Analyze▶Quality and Proces ▶Control Chart▶X Bar▶I menu vælg Process = Indhold af A ▶Sample Label = gruppe ▶Marker "Xbar, S, kSigma▶OK

Der fremkommer som i eksempel 9.1 kontrolkort, hvor 2 punkter falder udenfor kontrolgrænserne.
Som før fjernes disse, og man får følgende kontrol-kort.
Vi antager, at nu er processen i kontrol.



Det ses, at grænserne for middelværdien er
NKG = 0.25 og ØKG = 10.57

Hvis vi sammenligner med R-kortet var grænserne der
NKG = 0.26 og ØKG = 10.56
Forskellen skyldes, at spredningen jo bliver beregnet på en lidt forskellig måde, men forskellen er uden praktisk betydning.

9.5.2 Stikprøvestørrelse n = 1

I visse situationer er det praktisk umuligt at udtage mere end et element i stikprøven.
Det kan skyldes, at det tager så lang tid at få de nødvendige data, at det er urealistisk at undersøge mere end 1 element.
For at få et skøn for spredningen bruger man R-værdierne for 2 på hinanden følgende observationer (moving range) $MR_x = |x_i - x_{i-1}|$
Derfor kaldes kontrolkortet for spredningen for et "Moving range control chart"

Imidlertid er det naturligvis vigtigt, at fordelingen er rimelig normalfordelt, så man vil sædvanligvis først tegne et normalfordelingsplot.

Eksempel 9.3 Kontrolkort for enkeltobservationer

En olieraffineringsproces forløber så langsomt, at man må benytte en stikprøvestørrelse på n = 1.

Følgende observationer af oktantallet x for et bestemt produkt fandtes.

x	90.2	90.6	89.4	91.3	91.2	89.7	90.6	88.8	90.5	90.9	90.0	91.1	89.8	90.5	90.5	91.6	90.4	91.7

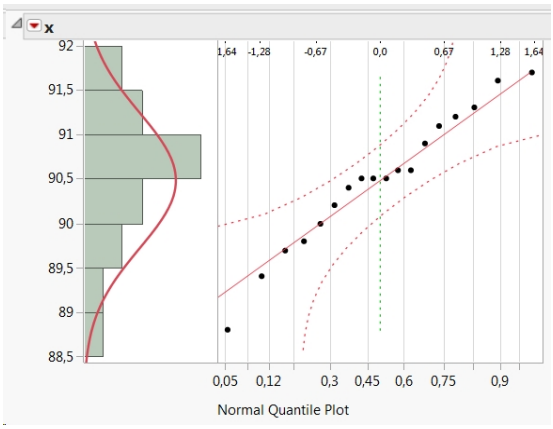
1) Vurder først om tallene er rimelig normalfordelt ved at tegne et normalfordelingsplot.
2) I det følgende antages, at fordelingen er rimelig normalfordelt.
Lav et "Moving Range kontrolkort" for processen.

Løsning.

1) Data indtastes i en søjle benævnt x.

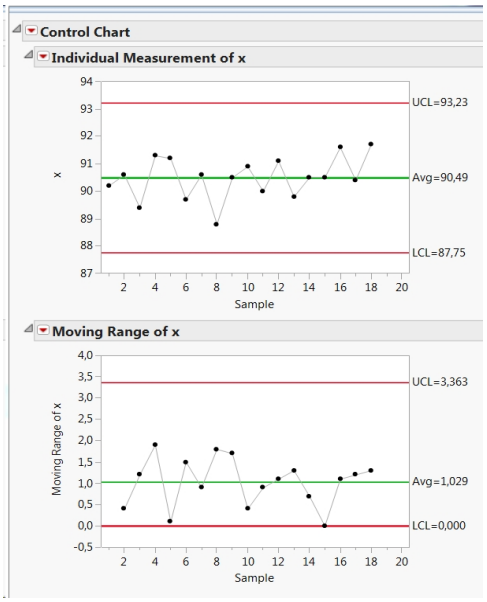
På værktøjslinien vælg Analyze ▶ Distribution
I den fremkomne menu dobbeltklikkes på "x" (under "Selected Columns") ▶ ok
Der fremkommer et "histogram", et "boxplot" og en række statistiske oplysninger.
For at kunne sammenligne med en normalfordeling tegnes en normalfordelingskurve:
Placer cursor på "rød pil ved x ▶ tryk på højre musetast og vælg "Continuous Fit ▶ normal
Der tegnes nu en normalfordelingskurve med samme middelværdi og spredning som fra data.
Cursor placeres på " rød pil ved x ▶ tryk på højre musetast og vælger "Normal Quantile Plot"
Der er nu bl.a. fremkommet følgende figur (har slettet markering ved Box plot)

Da punkterne ligger nogenlunde på en ret linie antages fordelingen at være rimelig normalfordelt



2) På værktøjslinien vælg Analyze ▶ Quality and Proces ▶ Control Chart ▶ IR ▶ indsæt x ▶ OK
Der fremkommer følgende to kontrolkort

Da alle punkter falder indenfor kontrolgrænserne er processen i kontrol.



Behandlingen af disse kort svarer til dem vi anvendte ved de øvrige kontrolkort.
Det skal dog nævnes, at disse kort har svært ved at opdage små ændringer i processens middelværdi. Den er også meget følsom overfor selv mindre afvigelser fra normalfordelingen.
Derfor anvendes ofte i stedet enten såkaldte "Cusum" kort eller "EWMA" kort (findes i SAS.JMP).

9.6. Løbende kontrol

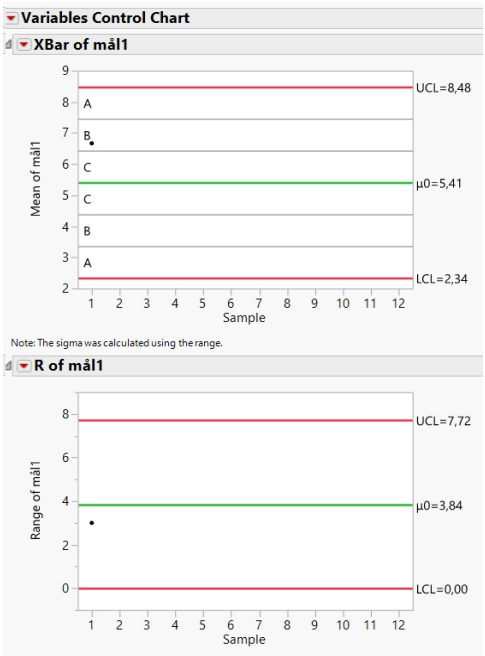
Når man har laver kontrolkortet skal det jo anvendes til at holde kontrol med at produktionen ikke af en eller anden grund begynder at ændre sig.
Det følgende eksempel viser hvordan.

Eksempel 9.4. Løbende kontrol.

Lad der være oprettet et kontrolkort med n = 3 , $\bar{\bar{x}} = 5.41$ og gennemsnittet af r = 3.84
Man kalder en søjle for "Målinger" og indtaster 3 tal eksempelvis 7 8 5.

Analyze ▶*Quality and Proces* ▶*Control Chart*▶*X Bar*▶*I menu vælg Process = målinger*▶ *Sample Size Constant* ▶*Vælg stikprøvestørelse (3)*▶*Specify Stats* ▶*indsæt mean(range) = 3,84 og mean(measure)= 5.41*▶*OK*

Der fremkommer nu nogle kontrolkort for xBar og R
Vælg "rød pil ved overskrift xBar
marker ▶tests▶ alle test og Show Center lines, Show Control lines, Show Zones, Test beyond limits.
Vi har nu følgende kort

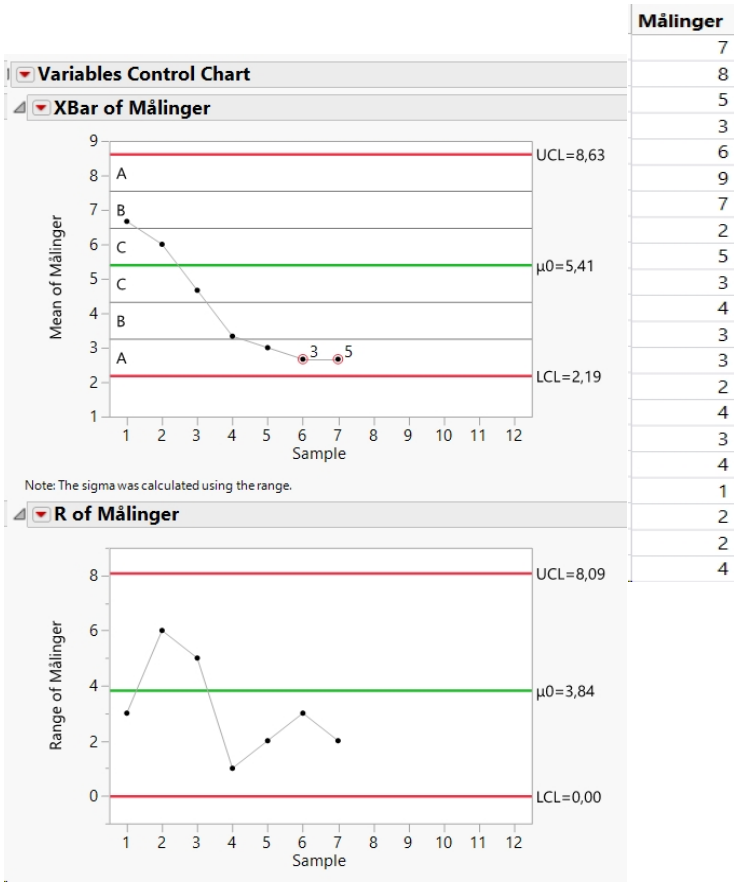


Vi kan nu løbende indtaste vore målinger.
Trykker man på Tests (i menuen fra før), ser man, at der er 8 test .
Markerer man eksempelvis Test 3 ses hvad den tester.

Test 1: 1 point beyond Zone A.
Test 2: 9 points in a row in a single (upper or lower) side of Zone C.
Test 3: 6 points in a row steadily increasing or decreasing.
Test 4: 14 points in a row alternating up and down.
Test 5: 2 out of 3 points in a row in Zone A or beyond.
Test 6: 4 out of 5 points in a row in Zone B or beyond.
Test 7: 15 points in a row in Zone C above and below the center line.
Test 8: 8 points in a row on both sides of the center line with none in Zones C.

På basis af nogle indtastede tal fremkommer følgende kontrolkort med indtegnede punkter, og en tydelig markering af hvilke alarmgrænser der er overtrådt.

Som det ses, er der slået alarm nr 3 og 5.



9.7 TOLERANCEGRÆNSER OG KAPABILITET

I forbindelse med en fabrikation er der ofte fastsat "tolerancegrænser" eller "specifikationsgrænser. Det kan enten være en nedre tolerancegrænse NTG_x og/eller en øvre tolerancegrænse $ØTG_x$,idet man ved fabrikationen ønsker/forlanger, for procesvariablen X , at $X \geq NTG_x$ henholdsvis $X \leq ØTG_x$.

Produktion, for hvilken X falder udenfor tolerancegrænsen/tolerancegrænserne, betragtes altså som fejlproduktion.

Bemærk, at der her ikke er tale om stikprøvens gennemsnit, men om den enkelte værdi af procesvariablen.

I det følgende forudsætter vi, at der er fastsat såvel en nedre som en øvre tolerancegrænse.

Ved **"kapabiliteten"** af en proces forstås processens evne til at producere inden for et specificeret toleranceinterval.

Som omtalt , er "det naturlige variationsområde" for en proces en variation på 6σ, nemlig fra -3σ til $+3\sigma$.

Når der ved en given fabrikation er fastsat et toleranceinterval $[NTG_x; ØTG_x]$ er det derfor nærliggende at sammenligne dette med intervallet for den naturlige variation $[-3\sigma;3\sigma]$.

Hertil udregnes et **kapbilitetsindeks** $C_p = \frac{ØTG_x - NTG_x}{6\sigma}$ (PCR = proces capability ratio= C_p)

C_p er et mål for processens evne til produktion indenfor toleranceintervallet.

Der foreligger herved i en kontrolsituation en af flere muligheder:
Hvis $C_p \approx 1$ dvs. $ØTG_x - NTG_x \approx 6\sigma$ skulle man tro, at dette ville være en proces med lille fejlproduktion, da højst 0.27% af produktionen vil falde udenfor.
Imidlertid er det med en stor løbende produktion svært hele tiden at holde middelværdien midt mellem de to tolerancegrænse.
Endvidere er fordelingen muligvis ikke helt normalfordelt, og endvidere kendes spredningen jo ikke eksakt men man udregner på basis af nogle målinger er et estimat s.
Man vil derfor sædvanligvis først sige, at man har en produktion med en lille fejlproduktion, hvis $C_p > 1.33$. Det afhænger naturligvis af produktionen og hvilke krav man stiller. hvor man sætter grænsen, og i den såkaldte "6 sigma proces" kræves et $C_p \geq 2 \Leftrightarrow ØTG_x - NTG_x \geq 12\sigma$

1) C_p stor (større end eksempelvis 1.33 eller 2)

   I dette tilfælde kan fejlproduktion (næsten) helt undgås, forudsat middelværdien ligger rimelig tæt ved midtpunktet af tolerancegrænserne $\frac{ØTG_x - NTG_x}{2}$.

2) C_p lille (under 1.33 eller 2)

   I dette tilfælde udøves sædvanlig processtyring, f.eks. med anvendelse af $\bar{x}$ - kort og *R* - kort.

   En (for stor) fejlproduktion kan ikke undgås, men den kan søges minimeret ved stram styring af processen, hvorved procesniveauet søges fastholdt på midterværdien $\frac{ØTG_x - NTG_x}{2}$.

   Eventuelt foretages en totalinspektion; eventuelt gennemføres en produktionsforbedring, hvorved processens spredning formindskes; eventuelt aftales nye tolerancegrænser og/eller en ny pris for det fremstillede produkt osv.

Tallet P = $\frac{100}{C_p}$ angiver hvor stor en procentdel af toleranceområdet der bruges af processen.

Er eksempelvis $C_p = 1$ bruger processen 100% af toleranceintervallet fra NTG til ØTG.

Som et mål for den aktuelle kapabalitet indføres et centreringsindeks (performanceindeks)

$$C_{pk} = \min\left(\frac{ØTG_x - \bar{\bar{x}}}{3\sigma}, \frac{\bar{\bar{x}} - NTG_x}{3\sigma}\right)$$

Hvis $C_p \approx C_{pk}$ er processen meget godt centreret, mens $C_{pk} < C_p$ viser, at dette ikke er tilfældet.
Der er andre forslag til kapabilitetsindeks, som er mere følsom over for den situation, at middelværdien μ for processen afviger fra midtpunktet T af tolerancegrænserne.

En af disse er $C_{pkm} = \frac{C_{pk}}{\sqrt{1+\left(\frac{\mu - T}{\sigma}\right)^2}}$

Ensidede kapabilitetsindeks

Hvis der kun er en øvre eller en nedre tolerancegrænse anvendes i stedet

$C_{pu} = \frac{ØTG - \mu}{3 \cdot \sigma}$ hvis der kun er angivet en øvre tolerancegrænse.

$C_{pl} = \frac{\mu - NTG}{3 \cdot \sigma}$ hvis der kun er angivet en nedre tolerancegrænse.

Eksempel 9.5. Beregning af kapabilitetsindeks

I eksempel 9.1 var ønsket at koncentrationen blev holdt indenfor tolerancegrænserne NTG = 0,5 og ØTG = 13.

Beregn kapabilitetsindeks C_p og vurder ud fra den og en beregning af fejlprocenten om processen har en rimelig lille fejlprocent.

Løsning

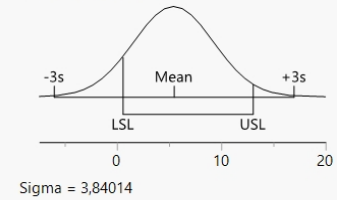
Data er allerede indlagt fra eksempel 9.1 med de 2 punkter udelukket, og vi har fundet de 2 kontrolkort. Vi vælger s-kortet (forskellen mellem beregning af s er dog meget lille, så vi kunne også godt vælge R-kort)

Vælg rød pil ved "Variable Control Chart" ▶ Capability ▶ Udfyld menu: "Lower spec. Limit" : 0,5 og "Upper Spec Limit" = 13 ▶ OK

Capability Analysis

Specification	Value	Portion	% Actual
Lower Spec Limit	0,5	Below LSL	12,1429
Spec Target	.	Above USL	2,8571
Upper Spec Limit	13	Total Outside	15,0000

Long Term Sigma

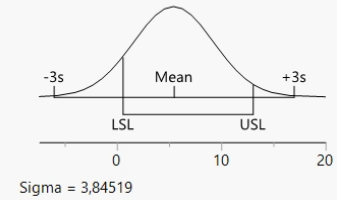


Sigma = 3,84014

Capability	Index	Lower CI	Upper CI
CP	0,543	0,479	0,606
CPK	0,426	0,351	0,500
CPM	.	.	.
CPL	0,426	0,351	0,500
CPU	0,659	0,564	0,754

Portion	Percent	PPM	Sigma Quality
Below LSL	10,0651	100650,52	2,778
Above USL	2,4008	24007,703	3,477
Total Outside	12,4658	124658,22	2,652

Control Chart Sigma



Sigma = 3,84519

Capability	Index
CP	0,542
CPK	0,425
CPM	.
CPL	0,425
CPU	0,658

Portion	Percent	PPM	Sigma Quality
Below LSL	10,0947	100946,68	2,776
Above USL	2,4155	24154,734	3,475
Total Outside	12,5101	125101,41	2,650

Man får $C_p = 0.543$ hvilket må siges at være meget lavt og da samtidig fejlprocenten på 12 5% er ret høj, giver processen alt for mange fejl, så den må ændres kraftigt.

Da C_{pk} er lidt lavere end C_p kunne man få en lidt lavere fejlprocent, hvis man kunne hæve processens gennemsnit til $\frac{13+5}{2} = 6.75$, men det vil næppe hjælpe meget.

Forklaring på forskellen mellem Long Term Sigma og Control Chart Sigma

Når man beregner kontrolkort beregnes spredningen ved at man for hver gruppe (på 5) beregner en spredning, og derefter tager gennemsnittet af tallene (30 værdier af s).

Ved "Long Term Sigma" beregner man spredningen af alle $30 \cdot 5 = 150$ tal på en gang.

Som man ser giver det en i dette tilfælde en meget lille forskel.

Imidlertid kan det for mindre stikprøver hvor målingerne måske ligger meget tæt sammen indenfor grupperne give en større forskel.

Man vælger derfor normalt at vælge udskrifterne ved "Long Term Sigma"

Eksempel 9.6. Beregning af kapabilitetsindeks (data ikke givet)

En proces var i statistisk kontrol med en estimeret procestilstand med middelværdi $\tilde{\mu} = 7.407$ og spredningen $\tilde{\sigma} = 2.840$

De udarbejdede kontrolkort kan derfor benyttes til den løbende kontrol.

Der bliver nu fastlagt følgende tolerancegrænser ØTG = 24 , NTG = 2

1) Beregn kapabilitetindexet C_p

   Forklar hvorfor dette tal ikke giver et sandt billede af processens evne til holde sig indenfor tolerancegrænserne.

2) Hvilken fejlprocent vil fabrikationen have?

3) Hvilket middelniveau bør processen ideelt søges indstillet og fastho1dt på.

Løsning:

1) $C_p = \dfrac{24-2}{6 \cdot 2.84} = 1.29$

   Selv om værdien er over 1 er der uden tvivl en ret stor fejlprocent, da estimatet for middelværdien 7.407 ligger ret langt gennemsnittet T = $\dfrac{24+2}{2} = 13$.

2) X = koncentrationen af stoffet ved en enkelt måling

   X antages normalfordelt med $\tilde{\mu} = 7.407$ og spredning $\tilde{\sigma} = 2.840$

   P(X>24)+P(X<2)=1-P(X<24)+P(X<2)= 1 - Normal Distribution(24, 7,407, 2,84) + Normal Distribution(2, 7,407, 2,84)

   =0.0285= 2.85%

3) Bør indstilles på en middelværdi på T = 13. ◆

Det kan være nyttigt at beregne et konfidensinterval for C_p

Konfidensinterval for C_p : $\tilde{C}_p \cdot \sqrt{\dfrac{\chi^2_{\alpha/2,n-1}}{n-1}} \le C_p \le \tilde{C}_p \cdot \sqrt{\dfrac{\chi^2_{1-\alpha/2,n-1}}{n-1}}$

Eksempel 9.7. Beregning af konfidensinterval for C_p

I forbindelse med en analyse af en proces fandt man, på basis af 20 målinger, at processens midtpunkt lå nogenlunde midt mellem øvre tolerancegrænse på 62 og nedre tolerancegrænse på 38. Endvidere skønnede man, at spredningen var ca. 1.75.

Find et 95% konfidensinterval for C_p

Løsning

n	øvt	nedt	s	cp	ned-kon	øvre-kon
20	62	38	1,75	2,2857142857	1,5649450045	3,0055794297

95% konfidensinterval for Cp $1.56 \le Cp \le 3.01$

Når man laver et sådan opstilling har man for fremtiden et "program" hvor men bare kan indsætte n, øvre og nedre tolerancegrænse

Formler:

$C_p =$ (øvt - nedt)/(6 * s) ned-kon= cp * sqrt(2/(n-1)) * ChiSquare Quantile(0,025, 19) øvre-kon= cp * sqrt(2/(n-1)) * ChiSquare Quantile(0,975, 19)

◆

9.8 PROCESVARIABLEN *X* ER DISKRET.

Vi vil her behandle den situation, hvor *X* enten er binomialfordelt eller Poissonfordelt.

9.8.1: *X* er binomialfordelt

Procesvariablen *X* er bestemt ved :

X = antal enheder med fejl ud af en stikprøve på *n*.

X er binomialfordelt *b*(*n*,*p*).

Sædvanligvis benyttes ved kontrol af fejlprocent ikke *p*, men middelværdien for binomialfordelingen $\mu = n \cdot p$ som parameter, og man siger man laver et *np* - kontrolkort.

Som beskrevet i forrige afsnit foretages en kontrolkortanalyse, ved at man opdeler i *k* undergrupper, som hver er karakteriseret ved en bestemt egenskab.

For hver undergruppe *i* på *n* enheder findes antallet af fejlenheder y_i.

Beregnes nu gennemsnittet $\bar{y} = \frac{1}{k}\sum_{i=1}^{k} y_i$, fås hermed et estimat for $\mu = n \cdot p$.

Et estimat for *p* er derfor $\hat{p} = \frac{\bar{y}}{n}$ og da spredningen for en binomialfordeling er $\sqrt{n \cdot p \cdot (1-p)}$

kan spredningen omformes til $\sqrt{\bar{y} \cdot \left(1 - \frac{\bar{y}}{n}\right)}$

Eksempelvis er $C_p = \frac{ØTG - NTG}{6 \cdot \sqrt{\bar{y} \cdot \left(1 - \frac{\bar{y}}{n}\right)}}$, ligesom det forklarer formlerne i oversigt 9.3.

Oversigt 9.3

np - kontrolkort. Procesvariablen *X* er binomialfordelt *b*(*n*,*p*)				
Forudsætning	Centerlinie	Nedre kontrolgrænse	Øvre kontrolgrænse	Estimater
Der udtages *k* undergrupper hver på *n* enheder. For hver undergruppe *i* findes antallet af fejlenheder y_i.	$\bar{y} = \frac{\sum_{i=1}^{k} y_i}{k}$	$\bar{y} - 3 \cdot \sqrt{\bar{y} \cdot \left(1 - \frac{\bar{y}}{n}\right)}$ (dog altid mindst 0)	$\bar{y} + 3 \cdot \sqrt{\bar{y} \cdot \left(1 - \frac{\bar{y}}{n}\right)}$	$\hat{p} \approx \frac{\bar{y}}{n}$

Eksempel 9.8 (*np* - kort)

En fabrikant af nogle specielle typer keramikfliser som er beregnet til at kunne klare høje temperaturer ønsker udarbejdet et kontrolkort.

a) Ved en løbende produktion af fliser udtoges 40 gange en stikprøve på 100 fliser. De blev undersøgt om de levede op til de forventede kvalitetsmål. Fliser der ikke opfyldte disse krav blev klassificeret som defekte
Resultatet var følgende:

Gruppe	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Antal defekte	8	6	4	4	3	7	3	6	9	5	7	2	6	11	4	6	7	4	9	6
Gruppe	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
Antal defekte	6	2	5	7	6	4	6	10	5	5	7	9	3	8	5	3	14	6	4	5

Lav et kontrolkort.

b) Man får at vide, at ØTG = 13 og NTG = 0.
Beregn kapabilitetindex C_p, og vurder om fejlprocenten er rimelig.

Løsning

a) Procesvariablen X er bestemt ved :

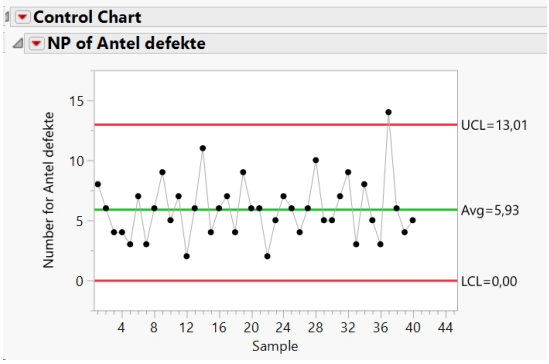
X = antal enheder med fejl af en produktion på 100 fliser

X er binomialfordelt $b(100, p)$

Data indtastes i en kolonne "antal defekte"

Vælg *Analyze* ▶ *Quality and Progress* ▶ *Control Chart* ▶ *NP* ▶ *I menu vælg Process = antal defekte* ▶ *Constant Size = 100* ▶ *OK*

Der fremkommer følgende kort



Da punkt 37, falder udenfor kontrolgrænserne, foretoges en nærmere undersøgelse af produktionsforholdene på det pågældende tidspunkt, men der blev herved ikke afsløret nogen tegn på væsentlige variationsårsager, jævnfør også, at punkt 37 ikke ligger meget over ØKG.

Man sluttede nu, at processen indtil videre kunne betragtes som værende i kontrol med en procestilstand på $\hat{p} = \frac{\bar{y}}{100} = 0.0593$.

b) $C_p = \frac{ØTG - NTG}{6 \cdot \sqrt{n \cdot p \cdot (1-p)}} = \frac{(13 - 0)}{\left(6 \cdot \sqrt[n]{5{,}93 \cdot \left(1 - \frac{5{,}93}{100}\right)}\right)} = \underline{0.917}$

Da værdien er mindre end 1.33, er processen ikke helt god. ◆

9.8.2: X er Poissonfordelt

Procesvariablen X er bestemt ved : X = antal fejl i en stikprøve på n enheder.

X antages Poissonfordelt $p(\mu)$

Middelværdien af Poissonfordelingen er $c = n \cdot \mu$

Som beskrevet i forrige afsnit foretages en kontrolkortanalyse, ved at man opdeler i k undergrupper, som hver er karakteriseret ved en bestemt egenskab.

For hver undergruppe i på n enheder findes antallet af fejlenheder c_i.

Beregnes nu gennemsnittet $\bar{c} = \frac{1}{k}\sum_{i=1}^{k} c_i$, fås hermed et estimat for middelværdien $c = n \cdot \mu$.

Vi har derfor, at et estimat for μ er $\hat{\mu} = \frac{\bar{c}}{n}$, og da spredningen for en Poissonfordeling er $\sqrt{\mu}$ forklarer dette formlerne i oversigt 9.4

Oversigt 9.4

c - kontrolkort. Procesvariablen X er Poissonfordelt $P(\mu)$				
Forudsætning	Centerlinie	Nedre kontrolgrænse	Øvre kontrolgrænse	Estimater
Der udtages k undergrupper hver på n enheder. For hver undergruppe i findes samlet antal fejl c_i.	$\bar{c} = \frac{\sum_{i=1}^{k} c_i}{k}$	$\bar{c} - 3 \cdot \sqrt{\bar{c}}$ (dog altid mindst 0)	$\bar{c} + 3 \cdot \sqrt{\bar{c}}$	$\hat{\mu} = \frac{\bar{c}}{n}$

Eksempel 9.9.(*c* - kort)

Ved en tekstilproduktion taltes anta1 fejl pr. 100 m^2 klæde. Følgende resultater fandtes (tidsmæssig rækkefølge for produktionen) :

nr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
antal fejl	3	3	6	3	0	1	3	5	8	7	4	10	5	5	5	4	2	4	5	1	2	0	1	1	4

Med henblik på en kontrolkortanalyse skal konstrueres et c-kort for processen

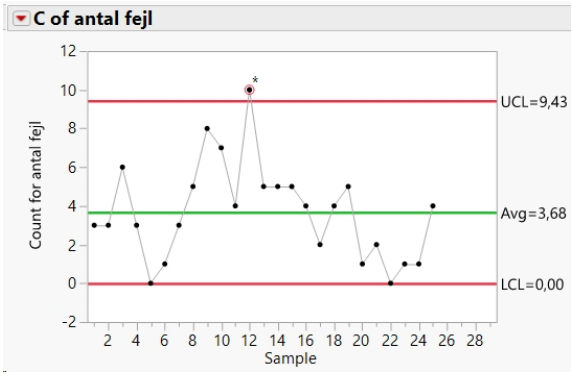
Løsning

Idet det antages, at den variable er Poissonfordelt dannes et c-kort på samme måde som np-kortet.

Data indtastes i en kolonne "antal fejl"

Vælg Analyze ▶Quality and Progress▶ Control Chart▶c ▶ I menu vælg Process = antal fejl ▶Constant Size =1 ▶OK

Vælg rød pil ved overskrift▶Test Beyond Limits

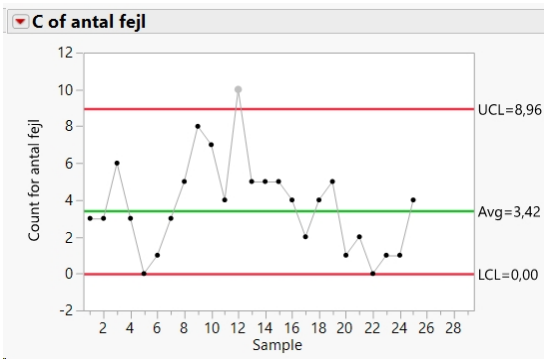


Punkt nr 12 falder uden for kontrolgrænserne, dvs. processen formodes at have været ude af statistisk kontrol på det pågældende tidspunkt.

Disse problemer bør undersøges nærmere.

Går man ud fra, at man vil se bort fra de målingen fjernes dette.

Fjernes den pågældende undergruppe, fås følgende reviderede kontrolgrænser ud fra den reviderede estimerede procestilstand:



Ingen af de resterende punkter falder uden for de reviderede kontrolgrænser.

$\bar{c}_1 = 3.42$, NKG = 0 , ØKG = 8.96 ◆

OPGAVER

Opgave 9.1

Ved en fabrikation af solbærsyltetøj tilstræbtes et gennemsnitligt nettoindhold af ca. 455 g solbærsyltetøj pr. glas. På glasetiketten anførtes: Nettoindhold 450 g.
Man havde på fabrikken mistanke om, at en af de automatiske fyldemaskiner havde svært ved at fastholde den ønskede nettovægt, og overvejede at foretage et hovedeftersyn af maskinen. For at afgøre, om et sådant burde foretages, udtog man med 5 minutters mellemrum i alt 18 gange 4 successivt producerede glas fra den omhandlede maskines produktion, og nettoindholdet af solbærsyltetøj bestemtes. Følgende resultater fandtes:

Stikprøve nr	1	2	3	4	5	6	7	8	9
Netto-indhold	452 455 454 453	452 457 458 456	458 455 458 455	456 458 456 456	460 456 454 457	457 457 452 456	452 450 454 453	455 456 456 457	454 459 456 459

Stikprøve nr	10	11	12	13	14	15	16	17	18
Netto-indhold	454 456 457 456	453 457 457 457	453 452 455 453	453 453 458 455	455 449 455 455	448 453 456 455	450 455 450 453	452 448 454 453	454 453 454 450

1) Lav et $\bar{x} - R$ - kontrolkort.
2) Estimér, hvor stor en del af produktionen der ville være fejlproduktion (underfyldte glas under 450 g), såfremt maskinen i statistisk kontrol med den i 1) konstaterede variation producerede med et middelindhold på 455g .

Opgave 9.2

Med henblik på indførelse af en proceskontrol for en løbende produktion er observeret følgende data:

Maskine nr	1	2	3	4	5	6	7	8	9	10
Data	12.9 11.6 13.0	l1.8 12.4 12.7	12.8 12.8 12.5	13.0 14.4 13.4	l1.9 12.3 12.5	12.1 12.7 12.0	13.6 12.0 12.4	13.8 l1.6 12.4	12.0 l1.4 l1.3	13.1 12.2 12.2

Foretag en kontrolkortanalyse med opstilling af $\bar{x}$ - kort og s - kort.
Punkter der falder udenfor kontrolgrænserne bortkastes

Opgave 9.3

Viskositeten af en polymer (stort kompliceret molekyle) måles en gang i timen.
Målinger foretaget de sidste 20 timer ses i følgende tabel.

test nr	1	2	3	4	5	6	7	8	9	10
Viskositet	2838	2785	3058	3064	2996	2882	2878	2920	3050	2870
test nr	11	12	13	14	15	16	17	18	19	20
Viskositet	3174	3102	2762	2975	2719	2861	2797	3078	2964	2805

1) Vurder først om tallene er rimelig normalfordelt ved at tegne et normalfordelingsplot.
I det følgende antages, at fordelingen er rimelig normalfordelt
2) Lav et "Moving Range kontrolkort" for processen.
3) Vurder om processen er i statistisk kontrol, og angiv et estimat for processens middelværdi og spredning.

Opgave 9.4

På en papirfabrik fabrikeredes en bestemt papirtype ved en af virksomhedens maskiner. Under fabrikationen måltes løbende værdien af en bestemt egenskab G (vægt/arealenhed).
Virksomheden besluttede sig i forbindelse med en eksportaftale til at indføre statistisk kvalitetskontrol og foretog indledningsvis en kontrolkortanalyse af G-værdier fra den omtalte maskine. Da man ikke havde nogen tekniske gruppeinddelingskriterier, lod man tidsmæssig nærhed være kriterium og valgte rationelle undergrupper på to på hinanden følgende observationer. Herved fremkom følgende gruppeinddeling:

Gruppe	1	2	3	4	5	6	7	8	9	10	11	12	13
G-værdi	86.2 85.4	85. 8 85. 6	85. 8 85. 8	85.7 86.4	86.6 86.1	86.1 86.0	85.6 86.0	86.0 85.9	86.1 86.3	85.3 85.5	85.6 85.2	85.6 85.5	86.0 85.3

Gruppe	14	15	16	17	18	19	20	21	22	23	24	25	26
G-værdi	85.3 86.1	85. 8 85. 9	85. 6 85. 6	85.5 85.2	85.3 85.8	85.3 85.6	85.8 86.1	86.3 86.3	86.0 85.8	86.0 85.6	86.3 86.2	86.0 85.8	85.9 86.3

1) Opstil R - $\bar{x}$ kontrolkort og vurder om processen efter eventuel fjernelse af enkelte punkter på kortene er i kontrol.
2) Vi antager i det følgende at processen er i kontrol med de fundne værdier for kontrolgrænser.
   Man benytter nu kontrolkortene til en løbende kontrol.
   Opstil et kontrolkort til den løbende kontrol (med alarmgrænser), og indsæt følgende punkter

Gruppe	1	2	3	4	5	6	7	8	9	10
G-værdi	86.6 86.1	86.3 86.7	86.6 86.5	86.5 86.6	86.2 86.3	85.9 85.3	85.4 85.1	85.2 85.1	85.6 85.0	86.9 85.3

Hvis en alarmgrænse forekommer, skal man angive hvilken alarmgrænse der er overtrådt.

Opgave 9.5

Ved en fabrikation af gips på basis af kalksten ønskedes indført en løbende proceskontrol med $\bar{x}$ og *s* - kort. Produktets kvalitet vurderedes bl.a. ved måling af DOP-absorptionen (g DOP/100 g pulver).

Ved den indledende kontrolkortanalyse benyttedes successive råvareleverancer som rationelle undergrupper, idet der af hver råvareleverance udtoges en stikprøve af størrelsen 5. Man fandt, at processen var i kontrol med middelværdi 21.94 og spredning 2.13.

Man opstiller på basis af disse målinger følgende $\bar{x}$ - s kontrolkort:

$NKG_s = 0$, $ØKG_s = 1.482$ $NKG_{\bar{x}} = 18.90$ $ØKG_{\bar{x}} = 24.98$.

Disse kontrolkort benyttes til den løbende kontrol.

1) Der er fastsat følgende tolerancegrænser for produktionen:
   NTG = 10.0 ØTG = 27.0
   Beregn kapabilitetsindeks C_p
   Forklar hvorfor dette tal ikke giver et sandt billede af processens evne til holde sig indenfor tolerancegrænserne.
2) Hvilken fejlprocent vil fabrikationen have?
3) Hvilket middelniveau bør processen ideelt søges indstillet og fastho1dt på.
   Beregn fejlprocenten hvis processen er indstillet på dette niveau og spredningen fortsat antages at være den i henhold til 1) estimerede?

Opgave 9.6

Ved indførelse af en proceskontrol observeret følgende data:

Gruppe	1	2	3	4	5	6	7	8	9	10
Data	13.0	l1.8	12.8	12.9	l1.9	12.1	13.6	13.8	12.0	13.1
	12.0	12.4	12.8	11.6	12.3	12.7	12.0	l1.6	l1.4	12.2
	13.4	12.7	12.5	13.0	12.5	12.0	12.4	12.4	l1.3	12.2

a) Foretag en kontrolkortanalyse med opstilling af $\bar{x}$ - kort og *R* - kort.
b) Der er fastlagt følgende tolerancegrænser for produktionen: NTG = 11 ØTG = 15
   Beregn kapabilitetsindeks $C_{p,}$ og vurder om procesen holder sig rimeligt indenfor grænserne.

Opgave 9.7

Fra en fyldeproces er udtaget 25 stikprøver af størrelsen 400. En enhed siges at være underfyldt, hvis enheden er påfyldt mindre end 95g. Antallet af underfyldte enheder i hver stikprøve er optalt. Følgende data fandtes:

Stikprøve nr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Antal underf. enh.	6	14	10	4	13	9	7	11	15	13	5	14	11	8	15	11	9	21	6	12	6	12	8	15	14

Foretag en kontrolkortanalyse og opstil kontrolkort, der kan benyttes til løbende kontrol.

Opgave 9.8
På en samlefabrik indgik en aksemontering. Som et led i en proceskontrol ønskede man specielt at undersøge risikoen for defekte samlinger. Der kan højst findes 1 defekt samling på en aksemontering.
Ved regelmæssig udtagning af stikprøver af størrelsen 250 fandtes følgende resultater:

Stikprøve nr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Antal defekte	11	12	6	18	10	8	12	10	7	11	11	7	5	8	7
Stikprøve nr	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Antal defekte	7	7	10	7	6	8	5	5	7	8	8	7	6	9	6

Lav kontrolkort med henblik på kontrol af den fremtidige produktion .

Opgave 9.9
Antallet af mindre fejl fundet ved den endelige kontrol af 22 DVD-pladespillere er vist i nedenstående tabel

Stikprøve nr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Antal fejl	1	1	3	7	8	10	5	13	0	19	14	6	9	11	15	8	3	6	7	4	9	10

a) Opstil et kontrolkort på basis af disse tal.
   Hvis der er punkter der falder udenfor kontrolkortet fjernes disse.
   Angiv kontrolgrænserne og gennemsnittet af fejl på basis af dette kort.
b) Idet tolerancegrænserne er ØTG = 18 og NTG = 0 skal man vurdere om antallet af mindre fejl er forsvindende.
   Vurderingen skal ske ved bl.a. at se på kapabilitetsindeks C_p .

Opgave 9.10
For antal defekte på kvadratiske ståloverflader skal foretages en løbende kontrol.
Antallet er poissonfordelt med en middelværdi på c = 4.
a) Opret et kontrolkort med alarmgrænser (bemærk, at der kun testes på 4 alarmer)
b) Indtast nedenstående tal på kortet, og angiv hvornår og hvilke alarmgrænser der er overtrådt.

nr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
antal	1	7	4	3	1	2	5	6	7	8	1	7	8	9	9	7	7	8	7	8	3	1	11	2	4	5

Opgave 9.11
Der foretages en løbende kontrol af plastikbægre på den måde, at man udtager 200 bægre og

optæller antallet af defekte bægre. Det oplyses, at ØTG = 14 og NTG = 0, samt at np=4.

a) Beregn kapabilitetindex C_p, og vurder om fejlprocenten er rimelig.
b) Lav et kontrolkort med passende alarmgrænser.

10. STATISTISK GODKENDELSESKONTROL

10.1 INDLEDNING.

Vi vil i dette kapitel betragte problemer af følgende type:

Eksempel 10.1. Problemstilling

Fra en **leverandør** til en **aftager** kommer varer i partier bestående af N emner. Hvert parti kan karakteriseres ved en ukendt procent af fejlemner , som kan variere fra parti til parti. Hvis denne fejlprocent er stor, vil aftageren ikke godkende partiet. Hvorledes afgøres, om et parti skal godkendes eller forkastes? ◆

I visse tilfælde er fejlene ved produktionen så uvæsentlige (måske kun skønhedsfejl) at aftagerne ud fra økonomiske overvejelser i en periode har foretrukket helt (eller næsten helt) at undlade en godkendelseskontrol, dvs. leverancerne accepteres uden inspektion.
Omvendt kan der være situationer, hvor konsekvenserne af godkendelse af fejlproduktion er så alvorlige, f.eks. helbredsmæssige konsekvenser for forbrugere af medicin (kritiske fejl), at l00%-kontrol principielt er nødvendig. Produktionen må tilrettelægges, så kritiske fejl ikke forekommer.

Sædvanligvis vil det imidlertid hverken være økonomisk rentabelt ikke at have nogen godkendelseskontrol (mange fejl giver mange klager og store erstatningsomkostninger) eller have en 100% - inspektion af hvert parti (inspektionsomkostningerne bliver for store).
Inspektion af alle emner i et parti stammende fra en massefabrikation af "ens" artikler vil erfaringsmæssigt alligevel sjældent giver 100% sikkerhed. Der hævdes at være erfaringer om, at en 100% inspektion af et meget stort antal emner kan være ned til kun 80% effektiv, dvs. at op til 20% af de defekte emner kan slippe igennem kontrollen ved en sådan 100% inspektion (på grund af kedsommeligt rutinearbejde). Hvis kontrollen er destruktiv (bevirker at emnet ødelægges), er en total inspektion naturligvis umulig.

Det mest økonomiske er sædvanligvis at foretage en stikprøvekontrol, udtage nogle emner fra partiet og på grundlag af en vurdering af disse at afgøre, om partiet skal godkendes eller forkastes. Dette medfører naturligvis, at uanset at kontrollen i gennemsnit fungerer godt, løber aftageren en vis risiko for, at et parti af dårlig kvalitet bliver accepteret, og leverandøren tilsvarende en risiko for, at et parti af god kvalitet bliver forkastet.
Indirekte kan godkendelseskontrollen i høj grad påvirke kvaliteten af en produktion gennem det pres, der lægges på producenten, om at forbedre kvaliteten af det fremstillede produkt, såfremt mange leverancer bliver forkastet ved kontrollen.
Hvis vurderingen af de udtagne emner baseres på en bedømmelse af hvert emne som fejlfrit eller defekt, taler man om "partikontrol ved alternativ variation".
Hvis vurderingen baseres på en måling af en kvantitativ egenskab ved hvert emne såsom diameteren af en aksel, , taler man om "partikontrol ved kontinuert variation".
Vi vil i det følgende kun gennemgå "partikontrol ved alternativ variation".
Med hensyn til kontrol med andre egenskaber, f.eks. middelværdi og spredning, henvises til egentlige bøger om kvalitetskontrol.

10.2. ENKELT-STIKPRØVEPLAN

Denne stikprøveplan er den mest benyttede i praksis på grund af, at den er så let at administrere. n kaldes **stikprøvestørrelsen**. c kaldes **godkendelsestallet**.

Definition af enkelt stikprøveplan.
Lad n være et positivt helt tal, og lad c være et ikke-negativt helt tal, hvor $c < n$.
Ved en enkelt - stikprøveplan (n,c) udtages tilfældigt en stikprøve af størrelsen n, og antallet x af defekte i stikprøven optælles.
Partiet godkendes, såfremt $x \leq c$ *; partiet forkastes, såfremt* $x > c$.

Eksempel 10.2. Enkelt stikprøveplan

Et legetøjsfirma modtager leverancer bestående af 10.000 billige dukker, og ønsker at kontrollere disses kvalitet ved stikprøveplanen $(n, c) = (100, 3)$.
Angiv hvorledes kontrollen skal foregå.
Løsning:
Af hver leverance (på 10.000 dukker) udtages en stikprøve på 100.
Hvis 3 eller færre af disse er defekte godkendes hele leverancen, ellers forkastes det.

◆

OC-kurve for en stikprøveplan
For en given stikprøveplan kan man beregne sandsynligheden for at et parti bliver godkendt (acceptsandsynligheden p_a) som funktion af partiets fejlprocent.
Grafen for denne funktion kaldes funktionens OC-kurve.

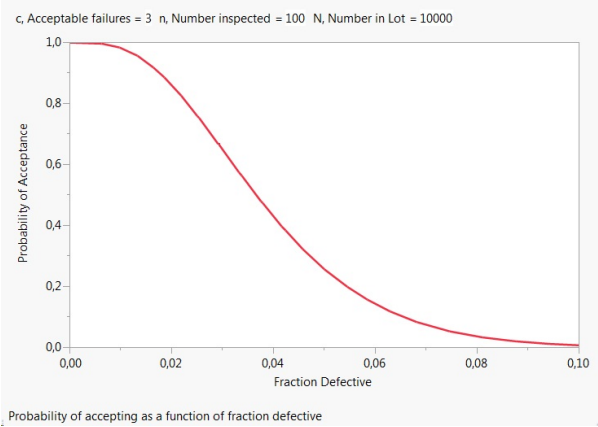
Beregning af acceptsandsynlighed.
Der udtages en stikprøve på n emner uden tilbagelægning ud af en leverance på N emner.
Lad c være godkendelsestallet.
X = antal fejlemner blandt n emner. Lad sandsynligheden for fejl i leverancen være p.
Acceptsandsynligheden er $p_a(p) = P(X \leq c)$.
Der er $M = N \cdot p$ fejlemner i partiet, dvs. X er hypergeometrisk fordelt h (N, M, n).
Sædvanligvis vil det gælde, at $\frac{n}{N} \leq \frac{1}{10}$, og derfor kan approksimeres med binomialfordelingen $b(n, p)$.

Eksempel 10.3. Tegning af OC - kurve.

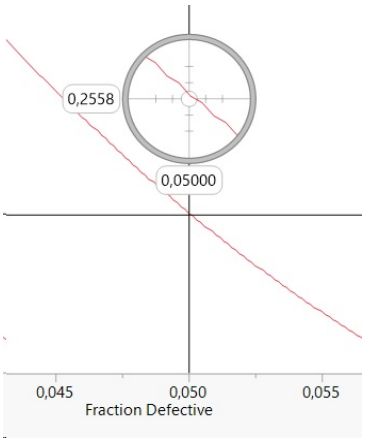
Et legetøjsfirma modtager leverancer bestående af $N = 10.000$ dukker, og ønsker at kontrollere disses kvalitet ved stikprøveplanen $(n, c) = (100, 3)$.
a) Tegn stikprøveplanens OC - kurve.
b) Aflæs på kurven acceptsandsynligheden for fejlprocent 5%.
c) Beregn acceptsandsynligheden p_a for fejlprocenten 5 %

Løsning:

a) Vælg View ▶ JMP-starter ▶ Control ▶ OC-curves ▶ Acceptance-single ▶ ok ▶ indsæt n = 100, c = 3, N = 10000



b) Vælg den smalle ramme øverst på figur .Der fremkommer en menu. Vælg Tools og crosshair (evt.magnifier)
Følgende figur fremkom(som man kan stabilisere ved at trykke på højremusetast), hvor man kan aflæse at for p = 5% skønnes aceptsandsynligheden at være 25.6%



c) X = antal fejlemner blandt n = 100 emner.
Er sandsynligheden for fejl p i leverancen er der $M = N \cdot p = 10000 \cdot p$ fejlemner i partiet.

X er derfor hypergeometrisk fordelt h(N, $N \cdot p$, n)= Hypergeometric Distribution(N, N * p, n, c)

Indtastes tallene i et skema
(hvorved det er hurtigt at beregne mange punkter), fås

N	n	c	p	accept
10000	100	3	0,05	0,2564...
10000	100	3	0,02	0,8598...
10000	100	3	0,06	0,1417...

$$P_a(0.05) = P(X \le 3) = \underline{0.256}$$

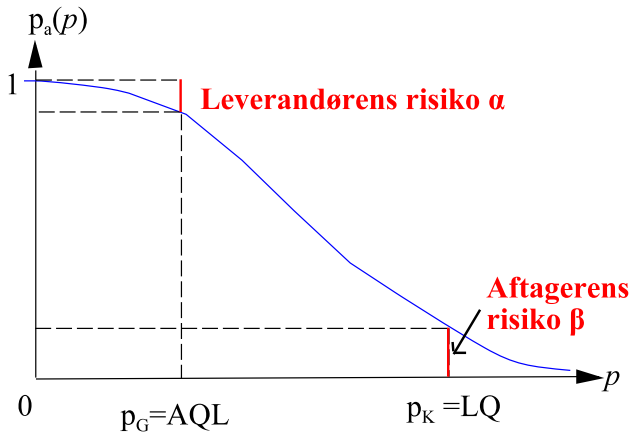
Ved at man udtager stikprøven fra en meget stor mængde, kunne man i stedet approksimere med binomialfordeling ◆

Valg af parametre i enkelt-stikprøveplan.

Til bestemmelse af stikprøvestørrelsen n og godkendelsestal c vælger de to parter, leverandør (producer) og aftager (consumer) to "risikopunkter" som på OC - kurven skal gå igennem.

1) den tilfredsstillende kvalitet AQL (Acceptable Quality Level), og den tilsvarende acceptsandsynlighed $P_a(AQL) = 1 - \alpha$, hvor α kaldes "leverandørens risiko".
(AQL,1 - α) kaldes leverandørens risikopunkt.
2) den utilfredsstillende kvalitet LQ (Limiting Quality), og den tilsvarende acceptsandsynlighed $P_a(LQ) = \beta$, hvor β kaldes "aftagerens risiko". (LQ, β) kaldes aftagerens risikopunkt.

De to risikopunkter må altid vælges ud fra de konkret foreliggende forhold (tekniske, økonomiske m.v.), idet *AQL* sættes lig en "lille" fejlandel p_G, og *LQ* sættes lig en noget større fejlandel p_K. (jævnfør den følgende figur)
Traditionelt vælges næsten altid $\alpha = 5\%$ og $\beta = 10\%$. Undertiden vælges dog $\beta = 5\%$.



En enkelt stikprøveplan er i princippet fastlagt ved disse specifikationer og kan approksimativt bestemmes ved hjælp af tabel 10.1.

Eksempel 10.4. Bestemmelse af stikprøveplan.
Ved levering af et partier på 10.000 dukker træffer en leverandør og en aftager aftale om at vælge en enkelt - stikprøveplan bestemt ved, at acceptsandsynligheden i tilfælde af 0.6% defekte dukker i partiet skal være 95%, og i tilfælde af 3,1% defekte skal være ca. 10%.
Bestem stikprøvestørrelsen *n* og godkendelsestallet *c*.
Løsning:
Tabel 10.1 anvendes

Idet AQL =0,6% og LQ = 3,1% er $R = \frac{LO}{AQL} = \frac{p_k}{p_G} = \frac{3.1}{0,6} = 5.1667$.

I tabel 10.1 i kolonnen for $\alpha = 5\%$ og $\beta = 10\%$ fås, at det største tal mindre end 5.17 er 4.89.

Dette svarer til, at godkendelsestallet $c = 3$. I kolonnen for $n \cdot p_g$ og $\alpha = 5\%$ findes

$$n \cdot p_G = 137 \Leftrightarrow n = \frac{137}{0.6} = 228.33 \approx 229$$

Vi finder derfor $\underline{c = 3 \text{ og } n = 229}$
Kontrol: Tallene indtastes i skemaet fra eksempel 10.3.

N	n	c	p	accept
10000	229	3	0,006	0,95175...
10000	229	3	0,031	0,07134...

p = 0.6% P(X<0.006)= 0.9518 = 95.18%

p = 3.1% P(X<0.031) = 0.07134 = 7.13%

Som det ses, passer det fint, hvis vi er leverandør, mens aftagerens risiko er noget mindre end 10%. Da vi jo kun kan udtage en stikprøve med et helt antal elementer, er det umuligt at få begge krav til at passe præcist.

Tabel 10.1. Bestemmelse af enkelt - stikprøveplaner

$R = \frac{p_k}{p_g}$ p_G og p_k regnes i procenter

c	α 1%			α 5%			α 10			β		
	β			β			β			1%	5%	10%
	1%	5%	10%	1%	5%	10%	1%	5%	10%	$n \cdot p_G$		
0	458	298	229	89.8	58.4	45.1	43.2	28.1	21.8	1.0	5.1	10.5
1	44.7	31.9	26.2	18.7	13.3	11.0	12.7	8.97	7.40	14.9	35.5	52.8
2	19.3	14.4	12.2	10.3	7.70	6.50	7.65	5.72	4.84	43.6	81.8	110
3	12.2	9.42	8.12	7.35	5.68	4.89	5.77	4.46	3.83	82.3	137	174
4	9.07	7.16	6.25	5.89	4.65	4.06	4.77	3.77	3.29	128	197	243
5	7.34	5.89	5.20	5.02	4.02	3.55	4.16	3.33	2.95	179	261	315
6	6.25	5.08	4.52	4.44	3.60	3.21	3.74	3.03	2.71	233	329	389
7	5.51	4.52	4.05	4.02	3.30	2.96	3.44	2.81	2.53	291	398	465
8	4.96	4.12	3.71	3.71	3.07	2.77	3.20	2.65	2.39	351	470	544
9	4.55	3.80	3.44	3.46	2.90	2.62	3.02	2.52	2.28	413	543	623
10	4.22	3.56	3.23	3.27	2.75	2.50	2.87	2.42	2.20	477	617	701
11	3.96	3.35	3.06	3.10	2.63	2.40	2.75	2.32	2.12	543	692	784
12	3.74	3.19	2.92	2.97	2.53	2.31	2.64	2.25	2.06	610	769	865
13	3.56	3.05	2.80	2.85	2.44	2.24	2.55	2.19	2.00	678	846	950
14	3.40	2.93	2.69	2.75	2.37	2.18	2.51	2.16	1.96	748	925	1030
15	3.27	2.82	2.60	2.67	2.30	2.12	2.41	2.08	1.91	818	1000	1110

10.3. REKTIFICERENDE KONTROL.

Undertiden aftales, at stikprøvekontrollen skal udføres på den måde, at leverandøren underkaster alle kasserede partier 100% - inspektion og "renser" disse partier, dvs. erstatter fejlemner med fejlfrie emner. En sådan kontrol kaldes for "rektificerende kontrol" (rensende kontrol).

En sådan aftale kan eksempelvis ske ved leverancer inden for samme virksomhed (intern kontrol) eller hvis en (dominerende) storaftager ikke er sikker på, at en leverandør kan overholde aftagerens krav til *AQL* - værdi.

I forbindelse med fastlæggelse af en enkelt stikprøveplan med rektificerende kontrol spiller følgende begreber en stor rolle.

***AOQ* (Average Qutgoing Quality):** Fejlandelen af de udgående partier når de indkommende partier produceres af en proces, der med en sandsynlighed på p frembringer fejlemner.

Der gælder formlen $AOQ = p \cdot P_a(p)$.

Bevis: Det modtagne parti har en "fejlsandsynlighed" på p , og sandsynligheden for at dette parti bliver godkendt er $P_a(p)$.

Det udgående parti (efter den rektificerende kontrol), vil derfor have en "fejlsandsynlighed" p_u på enten p eller på 0 afhængig af om det blev godkendt eller ej.

Vi får derfor: $AOQ = E(p_u) = p \cdot P_a(p) + 0 \cdot (1 - P_a(p) = p \cdot P_a(p)$ ◆

***AOQL* (Average Outgoing Quality Limit):** Den maksimale værdi af *AOQ*.

***ATI* (Average Total Inspection):** Den gennemsnitlige totale inspektionsstørrelse når de indkomne partier produceres af en proces, der med en sandsynlighed på p frembringer fejlemner.

Der gælder formlen $ATI = n + (N - n) \cdot (1 - P_a(p))$

Bevis:

Lader vi n_T betegne det totale antal inspicerede emner ved en rektificerende kontrol, må det gælde, at $n_T = \begin{cases} n & \text{med sandsynligheden } P_a(p) \\ N & \text{med sandsynligheden } 1 - P_a(p) \end{cases}$

Middelværdien bliver $E(n_T) = n \cdot P_a(p) + N \cdot (1 - P_a(p)) = n + (N - n) \cdot (1 - P_a(p))$

◆

Det følgende eksempel illustrerer beregningerne.

Eksempel 10.5 Rektificerende kontrol.

Vi betragter atter den i eksempel 10.4 omtalte situation, med kontrol af dukker, idet man har valgt enkelt-stikprøveplanen $(n,c) = (229, 3)$.

Vi antager nu, at der yderligere er aftalt, at kontrollen udføres som rektificerende kontrol.

Beregn acceptsandsynligheden AQL og ATI for p = 2%.

Løsning:

Da vi har et stort parti kan approksimeres med binomialfordelingen $b(229, 0.02)$

$P_a(0.02) = P(X \le 3)$ = Binomial Distribution(0.02,229, 3) = 0.326.

og dermed og $ATI = 229 + (10000 - 229) \cdot (1 - 0.326) = 6811.2$ ◆

10.4. DOBBELT STIKPRØVEPLAN

Anvendelsen af dobbelt-stikprøveplaner er administrativt noget besværligere end benyttelsen af enkelt-stikprøveplaner. Er det imidlertid meget dyrt (besværligt) at kontrollere emnerne, er det væsentligt at stikprøvestørrelsen bliver så lille som muligt og man kan i sådanne situationer se en fordel i at gå over til en dobbelt-stikprøveplan. Den gennemsnitlige stikprøvestørrelse vil derved ofte kunne formindskes i forhold til stikprøvestørrelsen n for en enkelt stikprøveplan (n,c), som går igennem de samme risikopunkter.

Ønskes en nærmere beregning af en sådan stikprøveplan henvises til bogen "Videregående statistik" side 32.

Opgaver

Opgave 10.1
Et firma modtager leverancer på 1000 stk kuglepenne. Leverancerne kontrolleres ved at man udtager 50 stk, og godkender partiet, hvis der er 2 eller færre defekte.
a) Tegn stikprøveplanens OC-kurve
b) Aflæs på kurven acceptsandsynligheden for en fejlprocent på 3%
c) Beregn accepsandsynligheden p_a for fejlprocenten 3%.

Opgave 10.2
En fabrik fremstiller billige disketter. Disketterne pakkes i kasser, som hver indeholder 100 disketter.
Produktionen er i statistisk kontrol, og det vides, at der i gennemsnit er 1 defekt diskette ud af 125 disketter .
1) Angiv øvre og nedre kontrolgrænse for et kontrolkort, baseret på, at man hver time udtager en kasse på 100 disketter, og optæller antallet af defekte disketter.

En forretningskæde med 5 forretninger køber hver måned et stort antal disketter fra fabrikken. Kæden træffer aftale med fabrikanten om at benytte en enke1t-stikprøveplan med følgende specifikationer:
Ved den tilfredsstillende kvalitet: fejlprocent 0.8, skal leverandørens risiko være 10%.
Ved den utilfredsstillende kvalitet: fejlprocent 3, skal aftagerens risiko være 10%.
2) Bestem stikprøvestørrelsen *n* (afrund til nærmeste med 100 delelige tal) og godkendelsestallet *c*.
3) Beregn aftagerens risiko ved en fejlprocent på 3.
4) Hver af de 5 forretninger foretager ovennævnte stikprøvekontrol.

Hvis produktionen er i statistisk kontrol således, at der som nævnt i gennemsnit forekommer 1 defekt diskette ud af 125 disketter, beregn da sandsynligheden for, at fabrikanten får et parti godkendt,
a) i 2 forretninger ud af de 5,
b) i alle 5 forretninger .

Opgave 10.3
Ved en statistisk partikontrol ved alternativ variation fastsættes en enkelt-stikprøveplan ved følgende specifikationer:
Ved den tilfredsstillende kvalitet: fejlprocent 3, skal leverandørens risiko være 5%.
Ved den utilfredsstillende kvalitet: fejlprocent 8, skal aftagerens risiko være 5%.
1) Bestem stikprøvestørrelsen *n* (afrund til nærmeste med 10 delelige tal) og godkendelsestallet c.
2) Tegn stikprøveplanens OC-kurve.
   Aflæs på tegningen acceptsandsynligheden for p = 4%
3) Beregn for den valgte stikprøveplan acceptsandsynligheden for p = 4%.

11 PLANLÆGNING AF FORSØG

11.1. INDLEDNING

Forsøg er en naturlig del af ingeniørmæssig og anden videnskabelig metode til at træffe beslutninger.
Antag eksempelvis, at en ingeniør skal studere virkningerne af 4 hærdningsmetoder på trykstyrken af et materiale.
Forsøget ville bestå i, at man fremstillede en række testmaterialer baseret på de 4 hærdningsmetoder, og derefter målte trykstyrken. På basis af disse data kunne man så anvende en statistisk metode til at finde den af de 4 metoder der i middel gav den største trykstyrke.

Alle forsøg er planlagte forsøg, men desværre er nogle forsøg særdeles dårlig planlagt, og resulterer i at kostbare ressourcer bliver benyttet ineffektivt.

Statistisk planlagte forsøg giver effektivitet og økonomi i den eksperimentelle proces, og brug af statistiske metoder i undersøgelsen af data resulterer i en **videnskabelig objektivitet** når man skal drage konklusioner.

Statistisk baserede forsøg er særlig nyttige til at forbedre en fremstillingsproces eller til at udvikle nye metoder.

Ved at benytte statistisk planlagte forsøg, kan ingeniøren bestemme hvilke af de mange procesvariable, såsom temperatur, tryk, hærdningsmetoder osv. der har den største betydning for udfaldet af processen.

Brugen af statistisk baserede forsøg kan derfor resultere i produkter, der er lettere at producere, produkter der har en bedre "performance" og stabilitet (mindre spredning) end konkurrenternes produkter, og kan blive udviklet og produceret på mindre tid, hvilket reducerer udviklingsomkostningerne.

11.2. NOMENKLATUR

I de følgende kapitler benyttes ord, som faktor, niveauer, behandlinger osv.
For at forstå hvad disse ord betyder, vil vi forklare dem ud fra følgende forsøg:

Eksempel 11.1 Nomenklatur

I forbindelse med nogle brudstyrkebestemmelser for Portland-cement udføres et fuldstændigt randomiseret forsøg til undersøgelse af middelbrudstyrkens afhængighed af cementblandere og cementknusere.
Med hver af 3 cementblandere udstøbtes efter blanding med vand 12 cementterninger, som efter en uges lagring underkastedes en brudstyrkeprøve ved hjælp af en af 4 cementknusere.
Forsøgsresultaterne var:

		Cementknusere											
		B_1			B_2			B_3			B_4		
Cementblandere	A_1	147	175	130	99	85	75	67	23	35	215	97	180
	A_2	211	145	163	131	100	145	75	45	71	151	157	167
	A_3	123	85	153	137	143	82	67	25	83	135	91	129

Beskriv forsøget.

Løsning:
Forsøget har to **faktorer**: Cementblander og Cementknuser.

Faktoren "Cementblander" har 3 **niveauer** A_1 ,A_2, A_3. (niveau hedder på engelsk" level")
Faktoren "Cementknuser" har 4 **niveauer** B_1 ,B_2, B_3., B_4

Forsøget har 12 **behandlinger** (engelsk treatment)
A_1B_1 , A_1B_2 , A_1B_3 , A_1B_4 , A_2B_1 , A_2B_2 , A_2B_3 , A_2B_4 ,A_3B_1 , A_3B_2 , A_3B_3 , A_3B_4
da der er 12 kombinationer af niveauerne (12 celler)

Hver behandling har 3 **gentagelser**, eksempelvis har behandlingen A_1B_1 3 **delforsøg,** der resulterede i **forsøgsresultaterne** 147 175 130

◆

Faktorer kan enten være **kvalitative** eller **kvantitative**.
En faktor som temperatur er kvantitativ, da den jo er en talvariabel, der kan antage alle mulige talværdier (indenfor et givet talområde).
En faktor som "Cementblander" i eksempel 11.1 er kvalitativ, da den kun har nogle fastlagte niveauer, og man ikke kan tale om eksempelvis cementblander 1.5.

11.3 KRAV TIL STATISTISK GYLDIGT FORSØG

For at nogle forsøgsresultater skal være statistisk gyldige, skal målingerne være statistisk **uafhængige** og være **repræsentative** for det man skal undersøge.
Ved statistisk uafhængighed forstås, at resultatet af et delforsøg ikke må afhænge af hvad der skete i de øvrige delforsøg.
Det er således ikke korrekt, hvis det arbejdshold, der foretager forsøgene først laver forsøgene med den ene cementblander- derved bliver dygtigere- og så laver forsøgene med de øvrige cementblandere.

Det er heller ikke korrekt, at man eksempelvis i eksempel 11.1 først havde målt holdbarheden af 10 blokke, - derefter foretager en test-opdager at man ikke kan vise signifikans.
Så tager man 10 blokke mere - testet på de 20 blokke osv., indtil man opnåede signifikans.
Dette er ikke "statistisk gyldige" forsøg.

Til belysning af hvad der er et "statistisk" gyldigt forsøg tages udgangspunkt i følgende eksempel.

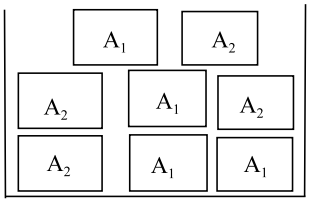
Eksempel 11.2. Planlægning af forsøg.
En fabrik der producerer maling, har udviklet to nye additiver A_1 og A_2, som bevirker en kortere tørretid. Additiv A_1 er det dyreste, men man forventer også, at det giver den korteste tørretid. På grund af prisforskellen, skal tørretiden dog være mindst 10 minutter kortere for A_1, før man vil gå over til den. For at undersøge disse forhold produceres nogle liter maling, som derefter deles op i mindre portioner. Til nogle af portionerne tilsættes additiv A_1 og til andre additiv A_2 . Tørretiden måles derefter. ◆

Generelt gælder, at hvert delforsøg i et forsøg udføres under en række forsøgsbetingelser.

Alle andre delforsøgsbetingelser end behandlingerne sammenfattes i et begreb, der kaldes forsøgsenheden.
I eksempel 11.2 er additiverne = behandlingerne og forsøgsenhederne er den enkelte portion maling, anvendt apparatur og personale, tidspunkt for delforsøget og de forhold med hensyn til temperatur, luftfugtighed osv., som gælder på forsøgstidspunktet. Bemærk, at forsøgsenhederne ofte indeholder faktorer, som ikke kan gøres ensartet fra delforsøg til delforsøg. Dette bevirker, at resultatet af de enkelte delforsøg varierer. Dette giver forsøgsvariablens variation" eller kort forsøgets "støj" .

Randomisering
For at sikre et statistisk gyldigt forsøg foretager man en såkaldt fuldstændig randomisering. Dette betyder at man ved lodtrækning fordeler "forsøgsenhederne" tilfældigt på behandlingerne. Dette sker, for at man ikke ubevidst kommer til at favorisere en af de to behandlinger. Hvis man eksempelvis helt systematisk i eksempel 11.2 først laver alle delforsøg med additiv A_1, kunne dette bevirke en favorisering af A_1 nemlig hvis forsøgsomstændighederne (apparater, personale, luftfugtighed) er mest "gunstige" ved begyndelsen af forsøgsperioden.
For at anskueliggøre denne randomiseringsproces antager vi, at vi i eksempel 11.2 skal lave 4 delforsøg med hver additiv. Endvidere antages, at delforsøgene skal indgå i den almindelige produktionsgang, dvs. at man af tidsmæssige, personalemæssige og på grund af en begrænset mængde apparatur må lade forsøgene forløbe over flere dage. Man tror ikke, at dage, apparatur og laborant har nogen væsentlig betydning for forsøgsresultaterne. Der er sandsynligvis også andre forhold udenfor vor kontrol, og som tilsammen bevirker, at selv om man udfører gentagne delforsøg med samme behandling, så får vi afvigende resultater.
For en sikkerheds skyld vælger vi imidlertid at randomisere dage, apparatur og laboranter
Lad os antage at der gælder følgende:
Mandag er det kun muligt at lave 1 delforsøg, idet apparatur nr. 1 og laborant A er de eneste der er ledige.
Tirsdag er der kapacitet ledig til 3 delforsøg:
Ét delforsøg hvor apparatur nr 2 og laborant A benyttes
Ét delforsøg hvor apparatur nr 1 og laborant B benyttes, og
Ét delforsøg hvor apparatur nr 3 og laborant C benyttes.
Onsdag kan der også laves 3 delforsøg osv. (se det følgende skema).

Forsøgsenheder			Behandlinger (additiv)
Dag	Apparatur	Laborant	
mandag	1	A	
tirsdag	2	A	
tirsdag	1	B	
tirsdag	3	C	
onsdag	3	B	
onsdag	4	C	
onsdag	1	A	
torsdag	3	B	



Vi foretager nu randomiseringen, som kort sagt er en form for lodtrækning. Sædvanligvis vil man benytte et program, der kan generere tilfældige tal (mange lommeregnere har et sådant program). For at anskueliggøre randomiseringen vil vi mere primitivt foretage lodtrækningen på følgende måde. På 4 sedler skrives A_1, på andre 4 sedler skrives A_2. Hver seddel krølles sammen til en kugle og placeres i en dåse. Sedlerne blandes ved at dåsen rystes (se figur). Hvis den første seddel der udtrækkes er A_2 så betyder det, at det delforsøg der mandag udføres med apparatur 1 og laborant A skal anvende additiv A_2 . Hvis den næste seddel der udtrækkes er A_1 så betyder det , at det delforsøg der tirsdag udføres med apparatur 2 og laborant A skal anvende additiv A_1 osv. Resultaterne kunne eksempelvis være som angivet på følgende skema:

Forsøgsenheder			Behandlinger (additiv)
Dag	Apparatur	Laborant	
mandag	1	A	A_2
tirsdag	2	A	A_1
tirsdag	1	B	A_1
tirsdag	3	C	A_2
onsdag	3	B	A_2
onsdag	4	C	A_1
onsdag	1	A	A_2
torsdag	3	B	A_1

På denne måde sikrer man sig, at vi får et så vidt muligt "statistisk gyldigt" forsøg. Hvis vi derfor efter beregninger (som ses i de følgende kapitler) konkluderer, at der er forskel på additiverne, så er det "korrekt", idet det ville være helt tilfældigt, hvis én af additiverne har været begunstiget med særlig gode forsøgsenheder.
Herved har man også sikret sig, at de to stikprøver (variable) er statistisk uafhængige.

Forsøg bør udføres, så alle behandlinger får lige mange gentagelser.
Ved planlægningen af forsøget er det ganske klart, at hvis man eksempelvis har ressourcer til at lave 20 delforsøg, så ville det være en meget dårlig plan, hvis man lavede 18 delforsøg med A_1 og kun 2 delforsøg med A_2 . Der bør i naturligvis tilstræbes at lave 10 delforsøg med hver behandling.
Delforsøg kan mislykkes, så målet i praksis ikke bliver opfyldt. I sådanne tilfælde kan de i de følgende kapitler anførte statistiske analyser dog stadig gennemføres. Testene bliver dog mindre "robuste" (dvs. mere afhængige af, at forudsætningerne gælder) , og beregningerne mere komplicerede.

Dimensionering
Man kan fristes til at tro, at jo flere gentagelser jo bedre.
Da spredningen på et gennemsnit er $\frac{s}{\sqrt{n}}$, er det klart, at hvis antal forsøg *n* er stort bliver spredningen lille, og så kan man finde, at der er en signifikant forskel selv om denne forskel er lille.
Imidlertid risikerer man med mange gentagelser at opdage så små forskelle, at de ikke har praktisk betydning, og så er de mange delforsøg jo spild af arbejdskraft og penge.
Endvidere gælder det jo, at hvis man laver 25 forsøg, så er spredningen formindsket med en faktor 5, mens hvis man laver 100 forsøg så er spredningen formindsket med en faktor 10. Der skal derfor særdeles mange forsøg til for alvor at formindske spredningen på gennemsnittet.

Analogt med forklaringen i kapitel 5 kan man under visse forudsætninger beregne hvor mange gentagelser (portioner) der skal anvendes for hver behandling, hvis $P(\text{fejl af type I}) \le \alpha$ og $P(\text{fejl af type II}) \le \beta$. Man skal naturligvis angive en bagatelgrænse Δ, men desuden kræver beregningerne, at spredningerne ved de to behandlinger er (tilnærmelsesvis) ens, og at man kan give et nogenlunde realistisk skøn for denne fælles spredning σ.
Det er naturligvis en svaghed ved dimensioneringen, at man inden forsøget er udført skal give et sådant skøn. En vurdering heraf kunne baseres på erfaringer fra tilsvarende forsøg. Findes sådanne erfaringer ikke må man først lave nogle få delforsøg og derfra få et rimeligt gæt på spredningen σ.
At spredningerne er nogenlunde ens vil i praksis ofte være tilfældet, da forsøgsenhederne jo er valgt ved randomisering. Når forsøget så er lavet, kan man (lidt sent) se, om man har skønnet rigtigt.

Dimensioneringen har kun betydning hvis man får en accept, da man så ved, at en eventuel forskel ikke har praktisk betydning. Hvis man får en forkastelse, så ved man der er en signifikant forskel, men om den er af praktisk betydning må en nærmere undersøgelse vise.

Formler for dimensionering af 2 variable findes i oversigten i kapitel 11 side 118

12 1 Faktor på to niveauer

12.1. NORMALFORDELTE VARIABLE

12.1.1. Test af differens mellem middelværdier

I dette afsnit benyttes et eksempel til at forklare metode, teststørrelse osv.
Derefter vises hvorledes det samme eksempel regnes med SAS.JMP.

Eksempel 12.1. Sammenligning af 2 normalfordelte variable

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.
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	90.1	92.4
M_2	92.4	94.6	93.0	94.0	92.4	92.9	96.4	92.1	92.8	93.4

Undersøg på basis af disse resultater, om det på et signifikansniveau på 5% kan påvises at tidsforbruget ved metode M_1 er mindst 2 minutter mindre end ved metode M_2.

Løsning:

a) Lad X_1 = tidsforbrug ved anvendelse af metode M_1 og
X_2 = tidsforbrug 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.

$H_0: \mu_2 - \mu_1 \le 2$ Begrundelse: Nulhypotesen udtrykker, at påstanden ikke er opfyldt (nul virkning).

$H: \mu_2 - \mu_1 > 2$ Begrundelse: Den alternative metode udtrykker det vi ønsker at bevise, så den angiver, at differensen i middeltidsforbruget er større end 2.

I beregningerne antager man, at $H_0: \mu_2 - \mu_1 = 2$, da en forkastelse så sikrer, at H gælder.

Der anvendes et færdigt program, der anvender en testmetode (Satterthwaites metode), som er robust overfor mindre afvigelser fra kravet om normalitet, når blot antallet af gentagelser er den samme.
Er det ikke tilfældet kan man stadig foretage testen, men så stilles der større krav til, at de variable X_1 og X_2 virkelig er normalfordelte.
Formlen for Satterthwaites metode kan findes i afsnit 12.5 oversigt 12.5.1.

Er den beregnede P-værdi < signifikansniveauet α forkastes H_0, dvs. vi har bevist den alternative hypotese H er sand. (sandsynligheden for vi dermed kommer med en forkert konklusion er mindre end α).
Man vil i et sådant tilfælde ofte være interesseret i hvor differensen ligger med 95% sikkerhed.
Man vil derfor se på et 95% konfidensinterval.
Er P-værdien > signifikansniveauet α accepters H_0, dvs. vi kan ikke på dette grundlag bevise, at H er sand.

Får man en accept og er P - værdien ikke meget større end signifikansniveauet α, så er det muligt at en "stærkere" *t* - test kunne give en forkastelse.
Denne stærkere t-test kræver imidlertid at de to spredninger kan antages at være ens.

Dette er ofte tilfældet på grund af randomiseringen, men er man i tvivl herom kan man først foretage en test af om spredningerne er ens.
Får man en accept heraf, har man naturligvis ikke hermed vist, at varianserne er ens, men da den følgende *t* - test af middelværdier er robust overfor mindre forskelle i varianserne, blot vi har samme antal gentagelser, er det tilladeligt i den følgende test af middelværdierne, at antage at varianserne er ens. ◆

Forklaring på formler
For hver af de 2 metoder udføres en række delforsøg. Lad antallet af forsøg være henholdsvis n_1 og n_2.
Vi antager, at X_1 og X_2 er statistisk uafhængige normalfordelte variable med henholdsvis middelværdierne μ_1 og μ_2 og spredningerne σ_1 og σ_2.
Nulhypotese $H_0: \mu_1 - \mu_2 = d$ (*d* er i eksemplet -2), eller $H_0: \mu_1 - \mu_2 - d = 0$,
Testproceduren baseres på fordelingen af differensen $Y = \overline{X}_1 - \overline{X}_2 - d$.
Ifølge additionssætningen (se evt. Statistik Grundbog) er Y normalfordelt med middelværdien $E(\overline{X}_1 - \overline{X}_2 - d) = E(\overline{X}_1) - E(\overline{X}_2) - d = \mu_1 - \mu_2 - d$
og $V(X_1 - X_2 - d) = V(X_1) + V(X_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.
Heraf følger, at $U = \frac{\overline{X}_1 - \overline{X}_2 - d}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ er normeret normalfordelt.
Teststørrelsen $U = \frac{\overline{X}_1 - \overline{X}_2 - d}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ gælder kun, hvis spredningerne σ_1 og σ_2 er kendt eksakt.
Kendes kun deres estimater s_1 og s_2 må der anvendes andre testprocedurer.
Får man ved en F-test en accept af, at varianserne er ens, pooles varianserne sammen til en fælles gennemsnitlig varians s_0^2 (se eventuelt oversigten i afsnit 12.5.1) og størrelsen $t = \frac{\overline{X}_1 - \overline{X}_2 - d}{\sqrt{\frac{s_0^2}{n_1} + \frac{s_0^2}{n_2}}} = \frac{\overline{X}_1 - \overline{X}_2 - d}{s_0\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ er nu *t* - fordelt.
Hvis stikprøvestørrelserne er store (begge over 30) er det dog tilstrækkelig nøjagtigt at anvende
$U = \frac{\overline{X}_1 - \overline{X}_2 - d}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ som teststørrelse.
I modsat fald kan anvendes en ret kompliceret procedure, der kaldes "Satterhwaite's metode".
Denne er beskrevet i afsnit 12.5.1. ◆

Eksempel 12.1. fortsat

Hypoteserne omskrives til

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

Data indtastes i 2 søjler. Derefter lægges tallet 2 til alle tal fra metode 1

Dette er her sket ved at man har lavet en ny søjle med ændringerne, og så ved hjælp af formula har skabt en ny søjle "ny tid"

Metode	Tidsforbrug
m1	87,8
m1	91,9
m1	89,8
m1	89
m1	92,6
m1	89,4
m1	91,4
m1	88,7
m1	90,1
m1	92,4
m2	92,4
m2	94,6
m2	93
m2	94
m2	92,4
m2	92,9
m2	96,4
m2	92,1
m2	92,8
m2	93,4

Metode	Tidsforbrug	x	ny tid
m1	87,8	2	89,8
m1	91,9	2	93,9
m1	89,8	2	91,8
m1	89	2	91
m1	92,6	2	94,6
m1	89,4	2	91,4
m1	91,4	2	93,4
m1	88,7	2	90,7
m1	90,1	2	92,1
m1	92,4	2	94,4
m2	92,4	0	92,4
m2	94,6	0	94,6
m2	93	0	93
m2	94	0	94
m2	92,4	0	92,4
m2	92,9	0	92,9
m2	96,4	0	96,4
m2	92,1	0	92,1
m2	92,8	0	92,8
m2	93,4	0	93,4

Vi har nu:

H_0: $m_1=m_2$ $\quad$ $H: m_1 < m_2$

Analyze►Fit y by x►Y-Response: ny tid ,x-Factor: Metode►ok►rød pil øverst på figur, t-test

t Test

m2-m1

Assuming unequal variances

Difference	1,0900t Ratio	1,628317
Std Err Dif	0,6694DF	16,99463
Upper CL Dif	2,5024Prob > |t|	0,1219
Lower CL Dif	-0,3224Prob > t	0,0609
Confidence	0,95Prob < t	0,9391

Da vi skal vise, at m_2-m_1>0 følger det heraf (se lige under t-Test) at vi skal se på Prob>t

Vi har følgelig, at P - værdi = 0.0609 .

Da P-værdi > 0.05 accepteres nulhypotesen, dvs. det er ikke muligt på dette grundlag at bevise, at tidsforbruget ved metode M_1 er 2 minutter mindre end ved metode M_2.

Dette fremgår også af konfidensintervallet, [-0.32 ; 2.50] som jo indeholder tallet 0.

Da P-værdien var så tæt ved 5%, vil vi nu forsøge med den stærkere t-test, hvor kravet er en fælles "poolet" spredning.Vi gør som før, men vælger nu Means/Anova/pooled t

Vi får som forventet en lidt mindre P-værdi = 0.0604, men det giver stadig en accept, så konklusionen er den samme.

Da P-værdien er så tæt ved 5% er der en god mulighed for, at tidsforbruget ved metode 1 faktisk er 2 minutter mindre end ved metode 2 (begår en fejl af type 2).

t Test

m2-m1

Assuming equal variances

Difference	1,0900	t Ratio	1,628317
Std Err Dif	0,6694	DF	18
Upper CL Dif	2,4964	Prob > |t|	0,1208
Lower CL Dif	-0,3164	Prob > t	0,0604
Confidence	0,95	Prob < t	0,9396

Det ville derfor være rimeligt at bede om at få foretaget flere forsøg.

Havde vi fået en P-værdi < 0.05, så ville næste træk være at teste om spredningerne var ens (se eventuelt eksempel 12.2 hvordan), da vi ellers ikke måtte have anvendt den sidste metode. ◆

Er de oprindelige tal ikke kendt, men kun gennemsnit og spredning må man anvende en t-test, hvor formlerne står i oversigt 12.5.1. Har man mange forsøg (over 30 for hver faktor) så er det beregningsmæssigt nemmere at anvende en z- test.

Eksempel 12.2a (oprindelige data ikke kendt)

En fabrik der producerer maling, har udviklet to nye additiver A_1 og A_2, som bevirker en kortere tørretid. Additiv A_1 er det dyreste, men man forventer også, at det giver den korteste tørretid. På grund af prisforskellen, skal tørretiden dog være mindst 10 minutter kortere for A_1, før man vil gå over til den. For at undersøge disse forhold produceres nogle liter maling, som derefter deles op i mindre portioner. Til nogle af portionerne tilsættes additiv A_1 og til andre additiv A_2 . Tørretiden måles derefter.
Man valgte at udføre 100 forsøg med hvert additiv.
Man fandt for A_1: $\bar{x}_1 = 118.6$ og $s_1 = 12.13$, og for A_2: $\bar{x}_2 = 133.2$ og $s_2 = 14.2$
Kan man ud fra disse data bevise på mindst signifikansniveau $\alpha = 0.05$, at malingen med additivet A_1 tilsat har en mindre middeltørretid end konkurrentens?
Løsning:
x_1 = tidsforbrug ved anvendelse af addtitiv A_1. Middelværdi μ_1
x_2 = tidsforbrug ved anvendelse af addtitiv A_2.Middelværdi μ_2
H_0: $\mu_1 + 10 = \mu_2$ H: $\mu_1 + 10 < \mu_2$ eller H_0: $\mu_1 = \mu_2 - 10$ H: $\mu_1 < \mu_2 - 10$
Da antal forsøg er over 30 kan benyttes z-test

s = $\sqrt{\frac{12{,}13^2 + 14{,}2^2}{100}}$ P-værdi= Normal Distribution(128,6 - 133,2, 0, s) = 0.00688 =0.69%

Da P-værdi < 5% forkastes H_0, dvs vi har bevist, at man skal anvende additiv A_2.

Eksempel 12.2b (oprindelige data ikke kendt)

Samme opgave som eksempel 12.2 a, men nu er antallet af forsøg ikke 100 men n_1= 30 og n_2= 29.
Da antallet ikke er over 30 må benyttes Satterthwaites metode som ses i oversigt 12.5.1.
Samme hypoteser som i eksempel 12.2 b. Bemærk at vi må sætte d = -10
Det ses, at spredningen er s = $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{12{,}13^2}{30} + \frac{14{,}2^2}{29}}$ = 3.443

t= $\frac{\bar{x}_1 - \bar{x}_2 - d}{s} = \frac{(128{,}6 - 133{,}2)}{s}$ =-1.336 p = t Distribution(t, 57) = 0.0934

P-værdi = 0.0934 = 9,34%
Da P-værdi 9,34% > 5% accepteres H_0 dvs. vi kan ikke bevise, at man skal anvende A_2.
Antal frihedsgrader er sat til (30-1) + (29-1) = 57
Hvis man bruger formlen for f fås 56.95 dvs. 57, så det samme.

s1	n1	s2	n2	a	b	c	g
12,13	30	14,2	29	4,9045...	6,95...	11,857...	56,949...

◆

Hvis nulhypotesen er H: $\mu_1 > \mu_2 + d$ må man beregne 1- den i programmet fundne P-værdi.

Test af Varians

Som det ses af eksempel 12.1 kan det sjældent blive nødvendig at benytte den lidt stærkere metode, da P-værdien nok bliver mindre, kun sjældent så meget, at det har betydning for konklusionen.
Imidlertid vil det i de næste kapitler være nødvendigt at teste om spredningen på 2 stikprøver er ens.Dette sker ved en såkaldt F - test hvor teststørrelsen er F = $\frac{s_1^2}{s_2^2}$

F-fordelingen er beskrevet i afsnit 12.5.2, hvor også fremgangsmåden ved testningen er angivet. Som det fremgår af eksempel 12.3 har SAS.JMP et program til testningen, så det er ikke nødvendigt at anvende teststørrelsen direkte.

Eksempel 12.3 . Test af varians

Samme problem som i eksempel 12.1
Undersøg ved en test på signifikansniveau på 5% om de to metoders varians er ens.

Løsning:

$H_0: \sigma_1^2 = \sigma_2^2$ mod $H: \sigma_1^2 \neq \sigma_2^2$

Analyze▶ Fit y by x▶Response:Tidsforbrug , Factor: Metode▶OK
Cursor på rød pil ▶Un-Equal Variance
Blandt en række udskrifter forekommer nedenstående
Tests that the Variances are Equal

Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
m1	10	1,668965	1,412000	1,370000
m2	10	1,302135	0,960000	0,880000

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	0,7193	1	18	0,4075
Brown-Forsythe	1,2610	1	18	0,2762
Levene	1,6453	1	18	0,2159
Bartlett	0,5199	1	.	0,4709
F Test 2-sided	1,6428	9	9	0,4711

I udskrift for F - test ses, at alle -værdier er større end 0.05.
Følgelig accepteres H_0, dvs. vi vil i en følgende test antager, at spredningerne er ens.
(selv om der kun var en enkelt der var over 0.05 ville man nok stadig antage, at spredningerne er ens) ◆

Dimensionering.

Eksempel 12.4 (fortsættelse af eksempel 10.2)

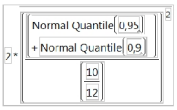
En fabrik der producerer maling, har udviklet to nye additiver A_1 og A_2, som bevirker en kortere tørretid. Additiv A_1 er det dyreste, men man forventer også, at det giver den korteste tørretid. På grund af prisforskellen, skal tørretiden dog være mindst 10 minutter kortere for A_1, før man vil gå over til den. For at undersøge disse forhold produceres nogle liter maling, som derefter deles op i mindre portioner. Til nogle af portionerne tilsættes additiv A_1 og til andre additiv A_2 . Tørretiden måles derefter.

a) Hvor mange delforsøg skal anvendes ved forsøget, hvis man ønsker, at $P(\text{fejl af type I}) = \alpha \leq 0.05$, $P(\text{fejl af type II}) = \beta \leq 0.10$ og bagatelgrænsen Δ =10 minutter, idet man fra mange tilsvarende forsøg ved, at den fælles spredning er σ = 12 minutter.
b) Samme spørgsmål og krav som i spørgsmål a), men nu antages, at man ikke kender spredningen, men ud fra nogle få delforsøg skønner, at den er ca. 12 minutter.

Løsning:

a) Af formlerne i oversigten i kapitel 12.5.1 fås $n \geq 2 \cdot \left(\frac{u_{1-\alpha} + u_{1-\beta}}{\frac{\Delta}{\sigma}} \right)^2 = 2 \cdot \left(\frac{u_{0,95} + u_{0,90}}{\frac{10}{12}} \right)^2 = 2 \cdot \left(\frac{1.645 + 1.282}{\frac{10}{12}} \right)^2$

SAS.JMP: Kald søjle for n ▶ Placer cursor i n's hoved ▶ højre musetast ▶ Formula ▶ vælg formler som vist



= 24.66 dvs. der skal udføres i alt $n = 25$ delforsøg af hver behandling

Spørgsmål b)
Her kræves løst en kompliceret ligning (se igen oversigten), så man vil sædvanligvis sige man benytter svaret i spørgsmål a) og hvis de er under 30 forsøg, så siger man blot det skal benyttes 30 forsøg, da man så kan antage spredninger er rimelig eksakt. ◆

12.1.2 Parvise observationer (blokforsøg)

Hvis observationerne i stikprøven udviser meget stor spredning, kan en test sandsynligvis ikke vise signifikans. Imidlertid er det måske så muligt at nedsætte spredningen ved at sammenligne observationerne to og to (i par). Man siger også, at man opdeler i blokke.
Det følgende eksempel illustrerer fremgangsmåden

Eksempel 12.5. Parvise observationer

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.
Da der er meget forskelligt hvor mange personer der går på trapperne i de forskellige bygninger (sammenlign blot sliddet på en skole og et plejehjem) vælger man at foretage et "blokforsøg" , med de 6 bygninger som 6 blokke.
I hver bygning lakeres hverandet trin (valgt ved lodtrækning) med lak A og resten med lak B.
Efter 3 måneders forløb måles graden af slid (i %) i hver bygning.
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.

Løsning

X_{gammel}= grad af slid ved anvendelse af sædvanlig lak B
X_{ny} = grad af slid ved anvendelse af ny lak A
Vi ser nu på differensen mellem sliddet i en bygning. (hvorved den store forskel mellem bygningerne elimineres)
Lad $D = X_{gammel} - X_{ny}$
D antages normalfordelt $n(\mu, \sigma)$, hvor såvel μ som σ er ukendte.

Da vi ønsker at teste om ny lak er mere slidstærk end gammel lak, dvs. den mest slidstærke lak slides mindst , bliver testen en ensidet t - test.
Nulhypotese $H_0 : \mu = 0$ Alternativ hypotese $H : \mu > 0$.
Data indtastes

Ny lak	Gammel lak
20,3	19,5
25,1	28,4
21,8	21,6
19,6	22
18,9	20,9
23,5	25,8

Vælg Analyze▶Specialized Modeling▶Matched Pairs▶Y:Paired response:Ny lak og Gammel lak ▶OK ▶Rød pil: Fjern markering ved "Plot Diff by Mean"
Der fremkommer følgende tabel:

Matched Pairs

Difference: Gammel lak-Ny lak

Gammel lak	23,0333	t-Ratio	2,268219
Ny lak	21,5333	DF	5
Mean Difference	1,5	Prob > |t|	0,0726
Std Error	0,66131	Prob > t	0,0363*
Upper 95%	3,19996	Prob < t	0,9637
Lower 95%	-0,2		
N	6		
Correlation	0,89502		

Da gammel lak - ny lak =D og H : $\mu > 0$ så ses på Prob>t

Man finder $\bar{x} = 1.5$ og P - værdi = $P(T > 1.5) = 0.0363$
Da P - værdi = 3.63% < 5% forkastes H_0 (svagt), dvs. den ny lak er mere slidstærk end den gamle.

12.2 BINOMIALFORDELTE VARIABLE.

Man kan ofte approksimere en binomialfordeling med en normalfordeling.
Det er en sådan approksimation, som formlerne i oversigten i afsnit 12.5.3 bygger på
Forudsætningen er:
X_1 og X_2 er **binomialfordelt** henholdsvis $b(n_1, p_1)$ og $b(n_2, p_2)$
Observerede stikprøveværdier x_1 og x_2.

Lad $\hat{p}_1 = \frac{x_1}{n_1}, \hat{p}_2 = \frac{x_2}{n_2}, \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$. **Forudsætning:** $n_1 \cdot \hat{p} \in [5; n_1 - 5], n_2 \cdot \hat{p} \in [5; n_2 - 5]$

I praksis vil disse forudsætninger for approksimation sædvanligvis være opfyldt.
Vi belyser beregningerne ved følgende eksempel.

Eksempel 12.6. Binomialfordelingstest.

Ved et forsøg der skulle afgøre om C - vitamin har en forebyggende virkning mod forkølelse, fik halvdelen af en gruppe på 280 franske skiløbere C - vitamin mens de øvrige fik kalktabletter (placebobehandling). Fordelingen skete randomiseret, og forsøgspersonerne var uvidende om gruppeinddeling og hvilket medikament de fik. Efter en passende tid optaltes hvor mange af forsøgspersonerne der var forkølede. Resultaterne kan ses af følgende skema:

	Forkølet	Ikke forkølet	Total
C-vitamin	17	122	139
Kalktabletter	31	109	140

Bemærk, at en enkelt forsøgsperson gled ud af forsøget, så grupperne blev ikke helt lige store.
1) Kan det på et signifikansniveau på 5% vises, at C - vitamin har en forebyggende virkning?
2) I bekræftende fald angiv er 95% konfidensinterval for differensen mellem parametrene.

Løsning:

X_1 = antal forkølede personer der har fået C-vitamin. X_1 er binomialfordelt b(139, p_1).

X_2 = antal forkølede personer der har fået Kalktabletter. X_2 er binomialfordelt b(140, p_2).

1) $H_0: p_1 - p_2 = 0$ mod $H: p_1 - p_2 < 0$ (da vi ønsker at vise, at $p_1 < p_2$).

Vi anvender nu formlerne i oversigt 12.5.3 side 120.

Vi ønsker at approksimere med normalfordeling.

$$\hat{p}_1 = \frac{x_1}{n_1} = \frac{17}{139},\ \hat{p}_2 = \frac{x_2}{n_2} = \frac{31}{140} \text{ og } \hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{17+31}{139+140} = \frac{48}{279}.$$

Da $n_1 \cdot \hat{p} = 139 \cdot \frac{48}{279} = 23.9 \in [5; 139-5]$ og $n_2 \cdot \hat{p} = 140 \cdot \frac{48}{279} = 24.1 \in [5; 140-5]$

er forudsætningerne for at approksimere med normalfordelingen opfyldt.

Den har middelværdi 0 og spredning s = $\sqrt{p \cdot (1-p) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{\frac{48}{279} \cdot \left(1 - \frac{48}{279}\right) \cdot \left(\frac{1}{139} + \frac{1}{140}\right)}$

P-værdi = Normal Distribution $\left(\frac{17}{139} - \frac{31}{140}, 0, \sqrt{\frac{48}{279} * \left(1 - \frac{48}{279}\right) * \left(\frac{1}{139} + \frac{1}{140}\right)}\right)$ =0.01414=1.41%

Da P - værdi = 0.0141 < 0.05 forkastes nulhypotesen (svagt)

Konklusion:

På signifikansniveau 5% er vist, at C-vitamin har en vis forebyggende virkning mod forkølelse,

2) 95% konfidensinterval;

Oversigterne er igen hentet fra oversigt 12.5.3

Radius r = Normal Quantile $(0{,}975) * \sqrt{\frac{\frac{17}{139} * \left(1 - \frac{17}{139}\right)}{139} + \frac{\frac{31}{140} * \left(1 - \frac{31}{140}\right)}{140}}$ = 0.08773

$p_1 - p_2 = \frac{17}{139} - \frac{31}{140}$ =-0.0991

Konfidensinterval for p_1-p_2 :[-0.0991-0.08773; -0.0991+0.08773]=[-0.187 ; -0.0111]

eller for p_2-p_1 [0.011;0,187]

12.3 POISSONFORDELTE VARIABLE.

Man kan ofte approksimere en Poisonfordeling med en normalfordeling.
Det er en sådan approksimation, som formlerne i oversigten i afsnit 12.5.4 bygger på. I praksis vil disse forudsætninger for approksimation sædvanligvis være opfyldt. Vi belyser anvendelsen af oversigten ved følgende eksempel.

Eksempel 12.7. Poissonfordelingstest.
En bestemt type TV-apparat produceres på 2 fabrikker A og B. Man har mistanke om, at der er forskel på antallet af loddefejl der findes i apparater fra de to fabrikker. For at teste dette, udtages af den løbende produktion stikprøver på 20 TV-apparater, og man optalte antallet af loddefejl i de 20 apparater. Resultaterne blev:
Fabrik A: På 20 apparater fandtes i alt 12 loddefejl
Fabrik B: På 19 apparater fandtes i alt 7 loddefejl (et apparat måtte udskydes)
Test på dette grundlag, om der er forskel på fejlintensiteten på de to fabrikker.
Løsning.
X_1 = antal loddefejl pr. apparat på fabrik A. X_1 antages Poissonfordelt $p(\mu_1)$.
X_2 = antal loddefejl pr. apparat på fabrik B. X_2 antages Poissonfordelt $p(\mu_2)$.
$H_0: \mu_1 = \mu_2$ mod $H: \mu_1 \neq \mu_2$ (da vi ønsker at vise, at $\mu_1 \neq \mu_2$)
Oversigten i kapitel 12.5.4 anvendes :

$$\bar{x}_1 = \frac{12}{20}, \bar{x}_2 = \frac{7}{19}, \text{ og } \quad \bar{x} = \frac{20 \cdot \frac{12}{20} + 19 \cdot \frac{7}{19}}{20+19} = \frac{12+7}{20+19} = \frac{19}{39}$$

Da $n_1 \cdot \bar{x} = 20 \cdot \frac{19}{39} = 9.74 \geq 5$ og $n_2 \cdot \bar{x} = 19 \cdot \frac{19}{39} = 9.26 \geq 5$ (lommeregner anvendt)
er forudsætningerne for at approksimere med normalfordelingen opfyldt.

$s = \sqrt{\frac{19}{39} \cdot \left(\frac{1}{20} + \frac{1}{19}\right)}$ og middelværdi 0

Da $\bar{x}_1 > \bar{x}_2$ er P-værdi = 1 - Normal Distribution $\left(\frac{12}{20} - \frac{7}{19}, 0, \sqrt{\frac{19}{39} * \left(\frac{1}{20} + \frac{1}{19}\right)}\right)$ =0.1502

Da P-værdi = 0.1502 > 0.025 accepteres nulhypotesen.
Konklusion: Man kan ikke på det grundlag vise, at der er forskel på fejlintensiteten på de to fabrikker, ◆

12.4. FORDELING UKENDT (Rangtest)

De testprocedurer vi har benyttet i de forrige kapitler har alle været baseret på, at i det mindste approksimativt kendte fordelingen (normal-, binomial- eller Poisson-fordelt) og testen vedrørte parametre i fordelingen såsom μ , σ eller p . Denne form for statistik kunne kaldes "parametrisk statistik". Kendes fordelingen ikke, og kan man heller ikke approksimere den til en kendt fordeling, så må man benytte de såkaldte "ikke- parametriske test". Disse forudsætter ikke, at fordelingen er kendt, og kunne derfor også kaldes "fordelingsfri test".
Det ligger udenfor denne bogs centrale emner, men der findes en beskrivelse heraf i den bog der ligger under min hjemmeside "Videregående statistik" kapitel 10.

12.5. OVERSIGT OVER CENTRALE FORMLER I KAPITEL 12

12.5.1. Test og konfidensinterval af differens $\mu_1 - \mu_2$ mellem middelværdier μ_1 og μ_2 for 2 normalfordelte variable .

X_1 og X_2 er **normalfordelte** henholdsvis $n(\mu_1, \sigma_1)$ og $n(\mu_2, \sigma_2)$.

Givet 2 stikprøver af X_1 og X_2. 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.

Satterthwaites test: Forudsætning σ_1 og σ_2 ukendte

Forkortelser: $a = \frac{s_1^2}{n_1}$, $b = \frac{s_2^2}{n_2}$, $c = a + b$, $g = \frac{c^2}{\frac{a^2}{n_1 - 1} + \frac{b^2}{n_2 - 1}}$

T er t - fordelt med frihedsgradstallet f, hvor f er det nærmeste hele tal, som er større end g.

Forudsætninger	Alternativ hypotese H	P - værdi	Beregning	H_0 forkastes
σ_1, σ_2 **ukendte** $t = \frac{\bar{x}_1 - \bar{x}_2 - d}{\sqrt{c}}$	$\mu_1 > \mu_2 + d$	$P(T \geq t)$	1 - t Distribution(t, f) eller t-test: se eksempel 12.1	P - værdi< α
	$\mu_1 < \mu_2 + d$	$P(T \leq t)$	t Distribution(t, f) eller t-test: se eksempel 12.1	
	$\mu_1 \neq \mu_2 + d$	$P(T \geq t)$ for $\bar{x}_1 > \bar{x}_2 + d$ $P(T \leq t)$ for $\bar{x}_1 < \bar{x}_2 + d$	som række 1 som række 2	P - værdi < $\frac{1}{2}\alpha$
	$100 \cdot (1-\alpha)\%$ **konfidensinterval for differens** $\mu_1 - \mu_2$: $\bar{x}_1 - \bar{x}_2 - tQuantile(1 - \frac{\alpha}{2}, f) \cdot \sqrt{c} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + tQuantile(1 - \frac{\alpha}{2}, f) \cdot \sqrt{c}$			

Z-test (eller U-test):Forudsætning: σ_1 og σ_2 kendt eksakt , eller n_1 og n_2 er større end 30

Forkortelser: $\bar{x} = \bar{x}_1 - \bar{x}_2 - d$ $\sigma = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Forudsætninger	Alternativ hypotese H	P - værdi	Beregning	H_0 forkastes
σ_1, σ_2 **kendte**	$\mu_1 > \mu_2 + d$	$P(Y \geq \bar{x})$	1-NormalDistribution $(\bar{x}, 0, \sigma)$	P - værdi< α
	$\mu_1 < \mu_2 + d$	$P(Y \leq \bar{x})$	NormalDistribution $(\bar{x}, 0, \sigma)$	
	$\mu_1 \neq \mu_2 + d$	$P(Y \geq \bar{x})$ for $\bar{x}_1 > \bar{x}_2 + d$ $P(Y \leq \bar{x})$ for $\bar{x}_1 < \bar{x}_2 + d$	som række 1 som række 2	P - værdi < $\frac{1}{2}\alpha$
	$100 \cdot (1-\alpha)\%$ **konfidensinterval for differens** $\mu_1 - \mu_2$: $\bar{x}_1 - \bar{x}_2 - Normal\ Quantil\left(1 - \frac{\alpha}{2}\right) \cdot \sigma \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + Normal\ Quantil\left(1 - \frac{\alpha}{2}\right) \cdot \sigma$			

Parvise observationer (blokforsøg)

t-Test: **Forudsætning: Forsøgene udføres "parvist"** (jævnfør eksempel 12.5).
Man danner differensen mellem tallene i hvert par (blok)
Lad D være denne differens, og lad D have middelværdi μ og spredning σ

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

Forudsætninger	Alternativ hypotese H	*P* - værdi	Beregning	H_0 forkastes
σ **ukendt.** $t = \frac{(\bar{x} - \mu_0)\cdot\sqrt{n}}{s}$.	$H: \mu > \mu_0$	$P(T \geq t)$	1 - t Distribution(t, f) eller Analyze, Matched pairs	*P* - værdi $< \alpha$
	$H: \mu < \mu_0$	$P(T \leq t)$	t Distribution(t, f) eller Analyze, Matched pairs	
	$H: \mu \neq \mu_0$	$P(T \geq t)$ for $\bar{x} > \mu_0$ $P(T \leq t)$ for $\bar{x} \leq \mu_0$	som række 1 som række 2	*P* - værdi $< \frac{1}{2}\alpha$ dog hvis t-test *P* - værdi $< \alpha$

Dimensionering:
$\Delta = |\mu_1 - \mu_2|$ er den mindste ændring der har praktisk interesse. α =P(type I fejl), β = *P*(type II fejl)

Spredninger ukendt, men antages at være nogenlunde ens og højst af størelsen σ

Ensidet test: $n \geq 2\cdot\left(\frac{u_{1-\alpha}+u_{1-\beta}}{\frac{\Delta}{\sigma}}\right)^2\cdot\left(\frac{t_{1-\alpha}(2n-2)}{u_{1-\alpha}}\right)^2$.

Tosidet test: I ovenstående formel erstattes α med $\frac{\alpha}{2}$

Spredninger kendt : Samme formel, idet man blot stryger sidste faktor med t- leddet.
Se eksempel 12.3.

12.5.2. Test af differens mellem varianser for to normalfordelte variable.

X_1 og X_2 er **normalfordelte** henholdsvis $n(\mu_1,\sigma_1)$ og $n(\mu_2,\sigma_2)$.

Givet 2 stikprøver af X_1 og X_2. Størrelse, gennemsnit og spredning henholdsvis n_1, $\bar{x}_1$, s_1 og n_2, $\bar{x}_2$, s_2.

Signifikansniveau er α.

Forkortelser: $F=\frac{s_1^2}{s_2^2}$, Q er F - fordelt $F(n_1-1,n_2-1)$

Forudsætninger	Alternativ hypotese H	P - værdi	Beregning	H_0 forkastes
μ_1,μ_2 **ukendte**	$H:\sigma_1^2>\sigma_2^2$	$P(Q\geq F)$	$FCdf(F,\infty,n_1-1,n_2-1)$ eller se eksempel 12.3	P - værdi$<\alpha$
	$H:\sigma_1^2<\sigma_2^2$	$P(Q\leq F)$	$FCdf(-\infty,F,n_1-1,n_2-1)$ eller se eksempel 12.3	
	$H:\sigma_1^2\neq\sigma_2^2$	$P(Q\geq F)$ for $F>1$ $P(Q\leq F)$ for F < 1	som række 1 som række 2	P - værdi $<\frac{1}{2}\alpha$ dog hvis Ftest P - værdi$<\alpha$
	$100\cdot(1-\alpha)\%$ **Konfidensinterval for forhold** $\frac{\sigma_1^2}{\sigma_2^2}$**:** $\frac{F}{F_{1-\frac{\alpha}{2}}(n_1-1,n_2-1)}\leq\frac{\sigma_1^2}{\sigma_2^2}\leq F\cdot F_{1-\frac{\alpha}{2}}(n_2-1,n_1-1)$			

μ_1 og μ_2 kendte: $F=\frac{\left((n_1-1)s_1^2+n_1\cdot(\bar{x}_1-\mu_1)^2\right)\cdot n_2}{\left((n_2-1)s_2^2+n_2\cdot(\bar{x}_2-\mu_2)^2\right)\cdot n_1}$, Q er F - fordelt $F(n_1,n_2)$.

μ_1 kendt og μ_2 ukendt: $F=\frac{(n_1-1)s_1^2+n_1\cdot(\bar{x}_1-\mu_1)^2}{s_2^2\cdot n_1}$, Q er F - fordelt $F(n_1,n_2-1)$

F - fordelingen

Til enhver F - fordeling er knyttet to hele positive tal f_T og f_N kaldet henholdsvis tællerfrihedsgradstallet og nævnerfrihedsgradstallet. En sådan fordeling benævnes $F(f_T,f_N)$.

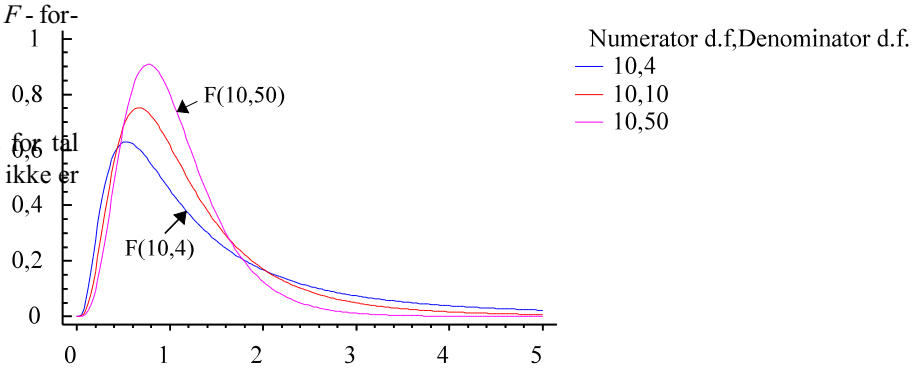
Det kan vises (se evt "afsnit til statistiske grundbegreber" 3C), at den variable $F=\frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}}$ er F - fordelt $F(f_1,f_2)$, hvor

f_1 og f_2 er de til s_1^2 og s_2^2 hørende frihedsgradstal.

Dette benyttes til testning af hypoteser vedrørende forholdet mellem to varianser.

På figuren er afbildet tæthedsfunktionen for F - fordelingerne
$F(10,4)$, $F(10,10)$ og F(10,50).

Det ses, at F fordelinger kun er defineret for tal større end eller lig nul, og at F-fordelinger ikke er symmetriske.



12.5.3. Test og konfidensinterval af p_1 - p_2 for 2 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.

Forkortelser: $\hat{p}_1 = \frac{x_1}{n_1}$, $\hat{p}_2 = \frac{x_2}{n_2}$, $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$, $s = \sqrt{\hat{p} \cdot (1 - \hat{p}) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

Y er normalfordelt variabel $n(0,s)$

Forudsætninger	Alternativ hypotese H	P - værdi	Beregning	H_0 forkastes
$n_1 \cdot \hat{p} \in [5; n_1 - 5]$ $n_2 \cdot \hat{p} \in [5; n_2 - 5]$	$H: p_1 > p_2$	$P(Y > \hat{p}_1 - \hat{p}_2)$	$1 - NormalDistribution(\hat{p}_1 - \hat{p}_2, 0, s)$	P - værdi $< \alpha$
	$H: p_1 < p_2$	$P(Y < \hat{p}_1 - \hat{p}_2)$	$NormalDistribution(\hat{p}_1 - \hat{p}_2, 0, s)$ se eksempel 12.6	
	$H: p_1 \neq p_2$	$P(Y > \hat{p}_1 - \hat{p}_2)$ for $\hat{p}_1 \geq \hat{p}_2$ $P(Y < \hat{p}_1 - \hat{p}_2)$ for $\hat{p}_1 < \hat{p}_2$	som række 1 som række 2	P - værdi $< \frac{1}{2}\alpha$
	$100 \cdot (1 - \alpha)\%$ **konfidensinterval for differens** $p_1 - p_2$ $\hat{p}_1 - \hat{p}_2 - NormalQuantile\left(1 - \frac{\alpha}{2}\right) \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \leq p_1 - p_2 \leq \hat{p}_1 - \hat{p}_2 + NormalQuantile\left(1 - \frac{\alpha}{2}\right) \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$			

12.5.4 Test og konfidensinterval af μ_1 - μ_2 for 2 Poissonfordelt variable

X_1 og X_2 er Poissonfordelte variable fordelt $p(\mu_1)$ og $p(\mu_2)$ hvor μ_1 og μ_2 er ukendte.
Der foreligger to stikprøver af størrelsen n_1 med gennemsnit $\bar{x}_1$ og n_2 med gennemsnit $\bar{x}_2$.
Signifikansniveau er α.

Forkortelser: $\bar{x} = \frac{n_1 \cdot \bar{x}_1 + n_2 \cdot \bar{x}_2}{n_1 + n_2}$ og $s = \sqrt{\bar{x}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

Y er normalfordelt variabel $n(0,s)$ se evt. eksempel 12.7

Forudsætninger	Alternativ hypotese H	P - værdi	Beregning	H_0 forkastes
$n_1 \cdot \bar{x} \geq 5$ $n_2 \cdot \bar{x} \geq 5$	$H: \mu_1 > \mu_2$	$P(Y > \bar{x}_1 - \bar{x}_2)$	$1 - NormalDistribution(\bar{x}_1 - \bar{x}_2, 0, s)$	P - værdi $< \alpha$
	$H: \mu_1 < \mu_2$	$P(Y < \bar{x}_1 - \bar{x}_2)$	$NormalDistribution(\bar{x}_1 - \bar{x}_2, 0, s)$	
	$H: \mu_1 \neq \mu_2$	$P(Y > \bar{x}_1 - \bar{x}_2)$ for $\bar{x}_1 \geq \bar{x}_2$ $P(Y < \bar{x}_1 - \bar{x}_2)$ for $\bar{x}_1 < \bar{x}_2$	som række 1 som række 2	P - værdi $< \frac{1}{2}\alpha$
	$100 \cdot (1 - \alpha)\%$ **konfidensinterval for differens** $\mu_1 - \mu_2$. $\bar{x}_1 - \bar{x}_2 - NormalQuantile\left(1 - \frac{\alpha}{2}\right) \cdot \sqrt{\frac{\bar{x}_1}{n_1} + \frac{\bar{x}_2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + NormalQuantile\left(1 - \frac{\alpha}{2}\right) \cdot \sqrt{\frac{\bar{x}_1}{n_1} + \frac{\bar{x}_2}{n_2}}$			

OPGAVER

Opgave 12.1

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 12.2

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) I tilfælde af signifikans beregn da et 95% konfidensinterval for differensen mellem de to gruppers middelværdier.

Opgave 12.3

I et laboratorium foretoges 15 uafhængige bestemmelser af furfurols kogepunkt, idet 8 af bestemmelserne foretoges af én kemiingeniør, de resterende bestemmelser af en anden kemiingeniør. Resultaterne var (^{0}C) :

1. ingeniør	162.2	161.3	161.9	161.2	163.4	162.4	162.5	162.0
2. ingeniør	163.3	162.6	161.8	163.8	163.0	163.2	164.1	

Undersøg, om de to ingeniørers resultater i middel er ens.

Opgave 12.4

Med henblik på at sammenligne de farmakologiske virkninger af stofferne morphin og nalbuphin foretoges et fuldstændigt randomiseret forsøg, hvorved man på 10 forsøgspersoner målte ændringen i pupildiameter (millimeter) efter indsprøjtning af en standarddosis af en opløsning af morphin (M) eller nalbuphin (N) .
Forsøgsplan og forsøgsresultater var:

M: 1.0	N: 0.0	M: 1.9	M: 2.0	N: 0.8	M: 0.8	M:0.1	N: - 0.3	N : 0.4	N: 0.2

a) Analyser forsøgsresultaterne.
b) Hvis der er en signifikant forskel skal der opstilles et 95%-konfidensinterval for differensen $\mu(M) - \mu(N)$ mellem de to middelværdier.

Opgave 12.5

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.
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.

a) Foretag en undersøgelse af, om valget var statistisk velmotiveret (bemærk: antal forsøg er forskellige) .
b) For en sikkerheds skyld skal man også kontrollere om varianserne er rimelig ens.
c) Hvis forslaget er velmotiveret skal der opstilles et 95% - konfidensinterval for differensen mellem middeludbytterne ved benyttelse af metoderne l og II.

Opgave 12.6

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 ved et fuldstændigt randomiseret forsøg tilfældigt n tråde af hver type og deres modstande måles.
Find det mindste antal n hvis man ønsker at $P(\text{fejl af type I}) = \alpha \le 0.05$, $P(\text{fejl af type II}) = \beta \le 0.05$ og bagatelgrænsen er $\Delta = 0.1$ ohm

1) og man ved, at spredningen $\sigma = 0.2$ ohm.
2) og man har en forhåndsformodning om, at spredningen er ca. $\sigma = 0.2$ ohm.
3) Hvilke konklusioner vedrørende behandlingernes virkning kan gøres, såfremt man ved testning af forsøgsresultaterne finder
   a) signifikans
   b) ingen signifikans
4) Hvilke yderligere analyser af forsøgsresultaterne bør foretages, såfremt testningen
   a) viser signifikans
   b) ikke viser signifikans.

Opgave 12.7

I et forsøg ønsker man at sammenligne udbyttet ved benyttelse af 2 reaktortyper.
Man ønsker at kunne påvise eventuelle forskelle i middeludbytte ned til ca. $\Delta = 6.0$.
Find den mindste værdi af n = "antal delforsøg med hver reaktortype", for hvilken
$P(\text{fejl af type I}) = \alpha \le 0.05$, $P(\text{fejl af type II}) = \beta \le 0.10$.
Man kender ikke spredningen eksakt, men mener, den højst er ca. 7 enheder.

Opgave 12.8

To sjællandske fabrikker producerer begge en bestemt type kvægfoder, for hvilken det ønskes, at proteinindholdet i færdigvaren skal være 26%. På de 2 fabrikkers driftslaboratorier foretoges følgende målinger af proteinindholdet i en uges produktion:

Fabrik 1	27.3	26.1	26.9	24.8	26.2	25.7	26.5

Fabrik 2	26.0	26.7	25.6	26.1	26.2	25.5	26.0	26.1	26.2	25.9

Foretag en statistisk vurdering af, om de to produktioner kan antages i middel at give kvægfoder med samme proteinindhold.

Opgave 12.9
Måling af intelligenskvotient på 16 tilfældigt udvalgte studerende ved en diplom-retning (med mere end 200 studerende) viste et gennemsnit på $\bar{x}_1$ = 107 og en empirisk varians på s_1^2 =100, medens en tilsvarende måling på 14 tilfældigt udvalgte studerende fra en anden diplomretning viste et gennemsnit på $\bar{x}_2$ =112 og en empirisk varians på s_2^2 = 64. Tyder disse tal på en forskel på studentermaterialet på de to retninger?

Opgave 12.10
I forbindelse med måling af luftforurening benyttes to apparater A og B til måling af mængden svovlmonooxid. Følgende målinger er foretaget daglig i en periode på 2 uger.

dag	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0.96	0.82	0.75	0.61	0.89	0.64	0.81	0.68	0.65	0.84	0.59	0.94	0.91	0.77
B	0.91	0.78	0.67	0.59	0.80	0.74	0.73	0.61	0.57	0.92	0.55	0.83	0.88	0.67

Man ønsker at undersøge om de to apparater i middel giver forskellige resultater på et signifikansniveau på 5%.

Opgave 12.11
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 12.12
To sjællandske fabrikker producerer begge en bestemt type kvægfoder, for hvilken det ønskes, at proteinindholdet i færdigvaren skal være 26%. For den omhandlede produktion er der fastsat en øvre og en nedre tolerancegrænse for proteinindholdet. Partier med et proteinindhold uden for toleranceintervallet klassificeres som "dumpere".
I en 3-måneders periode havde fabrik 1 af en produktion på 60 foderstofpartier 5 dumpere, medens fabrik 2 af en produktion på 100 foderstofpartier havde 12 dumpere. Kan det heraf statistisk konkluderes, at dumpeprocenten i middel har været størst for fabrik 2?

Opgave 12.13

1) Mange forbrugere tror, at såkaldte "mandagsbiler", dvs. biler produceret om mandagen, har flere alvorlige fejl (alvorlig fejl kan der kun være en af pr. bil) 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.
2) De 100 ovennævnte "mandagsbiler" havde i alt 1030 konstaterede større eller mindre enkeltfejl (enkeltfejl kan der godt være mange af på en bil), medens de 200 ovennævnte andre biler i alt havde 1899 konstaterede fejl. Tyder dette på, at der er forskel i fejlintensiteten på bilerne i de to grupper?

Opgave 12.14

To virksomheder A og B fremstiller dåser med nominelt 100 g rejeost. 10 tilfældigt udtagne dåser fra A's produktion og 20 tilfældigt udtagne dåser fra B's produktion viste fø1gende resultater:

	Virksomhed	
	A	B
Totalt antal rejer	81	216
Gennemsnittet $\bar{x}$ af nettoindhold	101.2 g	98.3 g
Empirisk spredning s af nettoindhold	1.0 g	2.7 g

1) Test, om det gennemsnitlige antal rejer pr. dåse kan antages at være det samme for virksomhedernes produktion.
2) Test, om det gennemsnitlige nettoindhold i en dåse kan antages at være det samme for virksomhedernes produktion.

Opgave 12.15

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.

13 1 FAKTOR PÅ MERE END 2 NIVEAUER

13.1 INDLEDNING

I kapitel 12 sammenlignede vi 2 middelværdier. I dette kapitel sammenlignes flere end to middelværdier. Det karakteristiske er, at de forekommende faktorer er kvalitative, dvs. har niveauer, som ikke er karakteriseret ved en målelig egenskab. Dette illustreres i det følgende eksempel.

Eksempel 13.1 (én faktor på 4 niveauer).
Virkningerne af 4 tilsætningsstoffer T_1, T_2, T_3, T_4 på mængden af urenheder ved en kemisk proces ønskes sammenlignet. For hvert tilsætningsstof måles mængden af "uønsket stof" 3 gange. Forsøgsresultaterne blev følgende:

Tilsætningsstof	T_1	T_2	T_3	T_4
Mængde uren-hed	108.1 110.2 111.7	105,1 110.3 108.6	115.8 110.7 113.0	117.1 119.2 111.8

Der ønskes fundet det tilsætningsstof der i middel giver den mindste urenhed. ◆

Faktoren "tilsætningsstof" siges at være en **kvalitativ** faktor på 4 **niveauer.**
Havde man eksempelvis i stedet på 4 tidspunkter målt mængden af uønsket stof'

Tid [i minutter fra starttidspunkt]	0	10	20	30
Mængde urenhed	108.11 110.2 111.7	105.1 110.3 108.6	115.8 110.7 113.0	117.1 119.2 111.8

siges faktoren "tid" at være en **kvantitativ** faktor.
En kvantitativ faktor er altså en "talfaktor", hvor det også har mening at spørge om mængde urenhed for mellemliggende værdier. Kvalitative faktorer er derimod ikke talbestemmende, og hvor det naturligvis ikke har mening at se på mellemliggende værdier (såsom tilsætningsstof nr. $T_{1.53}$).

Problemer, hvor faktorerne er kvalitative analyseres ved en "variansanalyse. Er faktorerne alle kvantitative, vil en metode kaldet "regressionsanalyse" være at foretrække.
Hvis nogle faktorer er kvantitative og nogle er kvalitative, kan man dog godt analysere problemet med variansanalyseteknikken, men da findes der mere effektive metoder, som dog ikke behandles i denne bog.

13.2 ENSIDET VARIANSANALYSE (normalfordelte variable)

13.2.1. Indledning

Vi vil i dette afsnit behandle problemer, af den type, som er vist i eksempel 13.1, dvs. med én faktor på mere end 2 niveauer.
Det vil i sådanne tilfælde være af interesse at teste om de til niveauerne svarende middelværdier afviger fra hinanden og i bekræftende fald hvilket niveau der giver den største/mindste værdi.
I eksempel 13.1 ønskes det således at finde det stof, der giver den mindste "middelurenhed".

Umiddelbart kunne man synes, at så foretager vi blot de samme parvise sammenligninger som i kapitel 12hvor vi så på differenser mellem 2 middelværdier. Problemet er imidlertid, at selv om de forskellige tilsætningsstofferne giver samme udbytte, så ville støjen i forsøget bevirke, at de mange gennemsnit fordeler sig "klokkeformet" (normalfordelt), og det vil sige, at den største og den mindste værdi let vil ligge så langt fra hinanden, at man ved at teste på deres differens fejlagtigt slutter, at der er forskel, selv om det faktisk ikke er tilfældet (fejl af type 1).
For at undgå dette, skal man derfor **altid** starte med at foretage den i det følgende beskrevne variansanlyse.
Giver den, at der ikke er signifikant forskel på middelværdierne, så skal man rette sig efter det, og ikke derefter begynde at se på konfidensintervaller.
Giver analysen, at der er en signifikant forskel, så ved man, at der i hvert fald er en signifikant forskel mellem det største og mindste middelværdi. Man kan så ved hjælp af passende konfidensintervaller forsøge at finde ud af om der også er en signifikant forskel mellem den største og næststørste værdi osv.

Som altid tilstræber man, at antallet af gentagelser af hvert niveau er det samme.
I notatet "Supplement til Anvendt statistik" som findes på min hjemmeside er angivet de generelle formler, som gælder, selv om eksempelvis et forsøg er mislykket, så man ikke har samme antal gentagelser.
SAS.JMP-programmerne anvender naturligvis disse generelle formler til beregningerne.

I afsnit 13.2.2 gives en forståelse for den teoretiske baggrund for variansanalyser.
Forklaringen understøttes af eksempel 13.1, hvor der vises, hvorledes man kan foretage beregningerne med en lommeregner eller PC, der kun har de sædvanlige sandsynlighedsfunktioner som normal- t- og F-fordeling, men ikke et egentligt statistikprogram.
Vi nøjes dog her med at se på hovedtilfældet, hvor antallet af gentagelser på hvert niveau er den samme.

13.2.2.Forklaring af metode og formler

Eksempel 13.1 (regnet uden brug af hjælpemidler)

Virkningerne af 4 tilsætningsstoffer T_1, T_2, T_3, T_4 på mængden af urenheder ved en kemisk proces ønskes sammenlignet. For hvert tilsætningsstof måles mængden af "uønsket stof" 3 gange. Forsøgsresultaterne blev følgende:

Tilsætningsstof	T_1	T_2	T_3	T_4
Mængde af urenhed	108.1 110.2 111.7	105.1 110.3 108.6	115.8 110.7 113.0	117.1 119.2 111.8

Find om muligt det tilsætningsstof der i middel giver den mindste urenhed og angiv i bekræftende fald et 95% konfidensinterval for middelværdien. ◆

Opstilling af nulhypotese.

Lad X_i = mængden af uønsket stof ved tilsætning af stof T_i hvor $i \in \{1,2,3,4\}$

Idet de 4 variables middelværdier kaldes μ_1, μ_2, μ_3 og μ_4 ønsker vi at teste nulhypotesen H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$, mod H: "mindst én middelværdierne er forskellig fra en af de øvrige".

Forsøgets udførelse. Forsøget skal udføres som et **fuldstændigt randomiseret forsøg**. (jævnfør kapitel 11vor et sådant forsøg er beskrevet). Derved sikrer vi, at der udføres et "statistisk gyldigt" forsøg. Hvis vi derfor, efter at have foretaget en ensidet variansanalyse, konkluderer, at der er forskel på tilsætningsstofferne, så er det "korrekt", idet det ville være helt tilfældigt, hvis én af tilsætningsstofferne har været begunstiget med særlig gode forsøgsenheder.

Beregning af gennemsnit og spredning. For at få et skøn for mængden af urenheder, udregnes gennemsnittene for hvert tilsætningsstof. Disse er angivet i nedenstående skema. Umiddelbart ud fra gennemsnit synes T_4 at adskille sig fra de tre øvrige, men hvis der er stor spredning, kan det måske blot være et tilfælde. Det er derfor naturligt at udregne spredningerne, hvilket derfor også er anført i skemaet.

	T_1	T_2	T_3	T_4
Gennemsnit	110.0	108.0	113.167	116.033
Spredning	1.8083	2.6514	2.5540	3.8136

Forudsætninger.

1) De 4 variable T_1, T_2, T_3 og T_4 skal være statistisk uafhængige.
   En måling af mængden af urenhed eksempelvis med tilsætningsstoffet T_3 må ikke afhænge af hvilke målinger der inden da er sket.
2) De 4 variable T_1, T_2, T_3 og T_4 skal være tilnærmelsesvis normalfordelte.
   Mindre afvigelser er tilladeligt. Man siger, at analysen er robust overfor afvigelser fra normalitet. Mest robust er den, hvis hvert niveau har samme antal gentagelser (som i eksempel 13.1, hvor alle har 3 gentagelser).

3) Der skal være varianshomogenitet.
Det betyder i eksempel 13.1, at målingerne af urenheder med tilsætningsstof T_1, skal have tilnærmelsesvis samme varians som målingerne af urenheder med T_2, T_3 og T_4.
Også her gælder det, at analysen er robust overfor mindre afvigelser, og mest robust, hvis antallet af gentagelser er det samme.

Sædvanligvis er forudsætningerne rimeligt opfyldt, hvis man har i forsøgsplanen har foretaget en randomisering.
Kun hvis man er i alvorlig tvivl vil man foretage en kontrol.

Kontrol af normalitet.
I notatet "Supplement til Anvendt statistik" som findes på min hjemmeside er vist hvorledes man kan lave et normalfordelingsplot af residualerne. Hvis disse ligger nogenlunde på en ret linie, er betingelsen opfyldt.

Kontrol af varianshomogenitet
Kan man få en accept af nulhypotesen
H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$ H: Mindst en varians er forskellig fra de øvrige
på et signifikansniveau på 5% (nogle mener 1% er ok hvis der er samme antal gentagelser) så er kravet rimeligt opfyldt.

Der findes forskellige forslag til hvorledes man tester dette. I notatet "Supplement til Anvendt statistik" som findes på min hjemmeside er angivet formlerne for 2 af disse, nemlig "Bartletts test" og "Levines test".
De er imidlertid så beregningsmæssigt kompliceret, at man nok altid vil benytte et egentlig statistikprogram som eksempelvis SAS.JMP til beregningerne.
Giver alle disse test en forkastelse, kan man eksempelvis prøve at tage logaritmen til alle tal i et forsøg på at nedsætte spredningen (vil ikke blive gennemgået her).
Er forudsætningerne ikke opfyldt kan man i stedet udføre en såkaldt "Kruskal Wallis""rangtest" (se evt. "Videregående statistik" kapitel 10 som findes på hjemmesiden larsen-net.dk) .
Når man ikke altid gør det, skyldes det at rangtest er svagere test, dvs. man vil ofte ikke opdage en signifikant forskel.

Pooling.
Da de 4 varianser antages at være nogenlunde ens, beregnes et vægtet gennemsnit, i forhold til frihedsgraderne (man foretager en "pooling").

$$s_{error}^2 = \frac{s_1^2 + s_2^2 + s_3^2 + s_4^2}{4} = \frac{\left(1{,}8083^2 + 2{,}6514^2 + 2{,}554^2 + 3{,}8136^2\right)}{4} = 7.8416$$

s_{error}^2 er variansen for "forsøgsfejlen" eller på engelsk "error".
s_{error}^2 har $N - r = 12 - 4 = 8$ frihedsgrader.
Det kan også ses af, at da hver af varianserne s_i^2 er baseret på $n = 3$ målinger har de hver 2 frihedsgrader ($f = n - 1 = 3 - 1$). s_{error}^2 har derfor $2 \cdot 4 = 8$ frihedsgrader.

F - test. Antages nulhypotesen H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ at være sand, dvs. udbyttet fra de 4 tilsætningsstoffer har samme middelværdi, er den eneste grund til, at vi ikke får samme gennemsnit i de 4 tilfælde, den ukontrollable "støj" (forsøgsvariablens variation) som forekommer ved forsøgets udførelse.

Indtastes de fire gennemsnit i eksempelvis SAS (Distribution) fås StdDev $s_{\bar{x}} = 3{,}5340$

Et gennemsnit af n tal har en varians, der er n gange mindre end variansen på den enkelte måling. I dette tilfælde er $n = 3$.

Et estimat for støjens varians forudsat nulhypotesen er sand er derfor

$s_R^2 = 3 \cdot s_{\bar{x}}^2 =$ $3 \cdot 3{,}534^2$ =37.467 Frihedsgradstallet er f_R =antal niveauer - 1 = 4 - 1 = 3.

Hvis nulhypotesen er sand burde $s_R^2 \approx s_{error}^2$ eller $\frac{s_R^2}{s_{error}^2} \approx 1$,mens hvis nulhypotesen er falsk(middel-

værdierne er forskellige) er $s_R^2 > s_{error}^2$ dvs. forholdet $F = \frac{s_R^2}{s_{error}^2}$ være signifikant større end 1.

Da $F = \frac{s_R^2}{s_{error}^2} =$ $\frac{37{,}467}{7{,}842}$ = 4,778 er spørgsmålet derfor, om dette tal er signifikant større end 1.

Da forholdet mellem de to varianser (som sædvanlig) er F - fordelt med $f_R = 3$ frihedsgrader i tælleren og $f_{error} = 8$ i nævneren kan vi afgøre dette ved at regne P - værdien ud.

P-værdi= P (F>4.7778) =1 - P(F<4.7778) = $1 - F\ Distribution(4{,}7778, 3, 8)$ =0.0342 = 3.42%

Da P - værdi = 0.034< 0.05 forkastes nulhypotesen (svagt).

Konklusion: De fire tilsætningsstoffer har ikke samme virkning.
Mindst 2 af middelurenhederne er forskellige.

Konfidensintervaller.

Disse beregnes **kun** hvis vi får en forkastelse af nulhypotesen, og dermed ved, at den største og den mindste middelværdi er signifikant forskellige. Om nogle af de øvrige middelværdier er "lige så gode" som den optimale vil ofte være af interesse.

Konfidensintervaller for middelværdi.

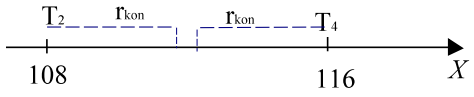
De "sædvanlige" konfidensintervaller for hvert niveau bestemmes ved (jævnfør eventuelt notatet "Supplement til Anvendt statistik" som findes på min hjemmeside):

$\bar{x}_i . \pm r_{kon}$ hvor $r_{kon} = t_{1-\frac{\alpha}{2}}(f_{error}) \cdot \sqrt{\frac{s_{error}^2}{n}}$, hvor n er det antal gentagelser der blev anvendt til beregning af gennemsnittet.

I vort tilfælde er $r_{kon} = t_{0.975}(8)\sqrt{\frac{7.842}{3}} =$ $t\ Quantile(0{,}975, 8) \cdot \sqrt{\frac{7{,}842}{3}}$ =3.728

og konfidensintervallet $\bar{x}_i . \pm r_{kon} = \bar{x}_i . \pm 3.728$

Konfidensintervallerne for 2 middelværdier overlapper derfor ikke, hvis afstanden er større end $2r_{kon}$.



Imidlertid vil disse intervaller være lidt for "brede", dvs. selv om der faktisk er en forskel på 2 middelværdier, så overlapper intervallerne måske hinanden en smule, så man ikke opdager det.
Idet $2 \cdot r_{kon} = 2 \cdot 3.728 = 7.456$ ses ikke overraskende, at der er forskel på T_2 og T_4 dvs. mellem største og mindste værdi.
Da vi ønske at finde det stof der giver mindst forurening sammenligne nu T_2 med T_1 og T_3.
Da afstandene er henholdsvis 2 og 5.17 kan de ikke adskilles.

LSD-intervaller (Least Significance Intervals)
Man kunne så foretage parvise sammenligninger svarende til de konfidensintervaller vi fandt i kapitel 12.

Problemet er imidlertid her, at hvis vi har *n* middelværdier, så vil der være $\frac{n(n-1)}{2}$ parvise sammenligninger. For hver af disse sammenligninger er der jo en vis sandsynlighed for at begå en fejl af type 1, dvs. påstå der er en forskel som reelt ikke er der. Sådanne fejl vil jo hobe sig op, hvis man foretager mange sammenligninger, så sandsynligheden for at begå en fejl af type 1 kunne blive betragtelig.

LSD - konfidensintervaller, der beror på parvise sammenligninger er bestemt ved, at deres radius er $r_{lsd} = r_{kon}\frac{\sqrt{2}}{2} = \frac{r_{kon}}{\sqrt{2}}$

Bevis:
Lad $Y = \bar{X}_1 - \bar{X}_2$

Vi har da (af kvadratsætning) at variansen $V(Y) = V(\bar{X}_1) + V(\bar{X}_2) = \frac{s_0^2}{n} + \frac{s_0^2}{n} = \frac{2 \cdot s_0^2}{n}$

Heraf følger, at Y har konfidensintervallet $\bar{x}_1 - \bar{x}_2 \pm r_1$ hvor $r_1 = t_{0.975}(f_0)\sqrt{\frac{2 \cdot s_0^2}{n}}$

Et LSD-interval omkring et enkelt punkt er følgelig $r_{LSD} = \frac{r_1}{2}$

Heraf fås $r_{LSD} = \frac{t_{0.975}(f_0)}{2}\sqrt{\frac{2 \cdot s_0^2}{n}} = \frac{t_{0.975}(f_0)}{2}\sqrt{2}\sqrt{\frac{s_0^2}{n}} = \frac{t_{0.975}(f_0)}{\sqrt{2}}\sqrt{\frac{s_0^2}{n}} = \frac{r_{kon}}{\sqrt{2}}$ ◆

Her vil man kunne opdage en forskel, hvis middelværdiernes afstand er større end $2 \cdot r_{lsd} = 2 \cdot \frac{r_{kon}}{\sqrt{2}} = \sqrt{2} \cdot r_{kon}$.

Da $\sqrt{2} \cdot 3.728 = 5.272$ ses nu, at T_2 giver mindre forurening end T_3 mens T_1 og T_2 stadig ikke kan adskilles.

De fleste statistikprogrammer har en række andre metoder til beregning af konfidensintervaller, som søger at formindske sandsynligheden for at begå fejl af type 1 og type 2.
Vi vil i dette notat kun se på ovennævnte to typer, og hvis vi har få middelværdier stole mest på

LSD-intervallerne.
Konklusion: LSD-intervallerne viser, at man får den mindste urenhed, hvis man vælger enten T_1 eller T_2 (de kan ikke adskilles).

Et 95% konfidensinterval for T_2 er [108- 3.728; 108+3.728] =[104.3 ; 111.7]

13.2.3. Beregning af ensidet variansanlyse

Eksempel 13.2 (eksempel 13.1 fortsat)

Virkningerne af 4 tilsætningsstoffer T_1, T_2, T_3, T_4 på mængden af urenheder ved en kemisk proces ønskes sammenlignet. For hvert tilsætningsstof måles mængden af "uønsket stof" 3 gange. Forsøgsresultaterne blev følgende:

Tilsætningsstof	T_1	T_2	T_3	T_4
Mængde urenhed	108.1 110.2 111.7	105.1 110.3 108.6	115.8 110.7 113.0	117.1 119.2 111.8

a) Kontroller om betingelserne om normalitet og varianshomogenitet er rimelig opfyldt.
b) Test på signifikansniveau på 5% om der er forskel på middelværdierne for de 4 tilsætningsstoffer
c) Find om muligt det tilsætningsstof der i middel giver den mindste urenhed og angiv i bekræftende fald et estimat for den mindste urenhed og et 95% konfidensinterval for middelværdien.

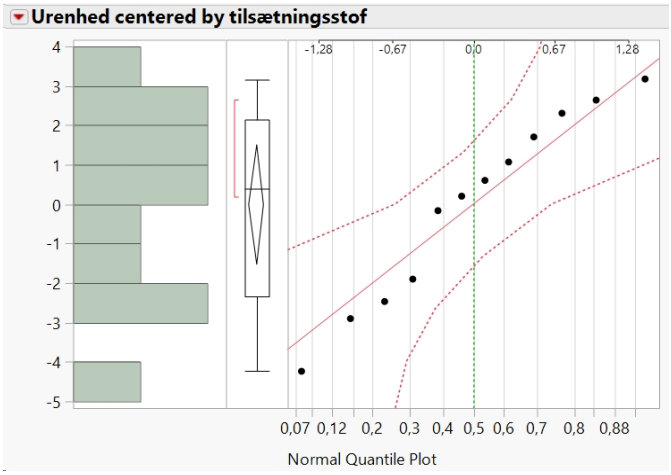
Løsning:
Data indtastes, "tilsætningsstof" skal være "rød" (se under Columns) og urenhed skal være "blå".

```
tilsætningsstof  urenhed
t1               108
t1               110
t1               112
t2               105
t2               110
t2               109
t3               116
osv
```

a) **Kontrol af forudsætninger**:
Selv om det er muligt i SAS.JMP at kontrollere forudsætningen om normalfordeling, behøver man ikke nødvendigvis at gøre det, da testen er robust overfor afvigelser.

Normalfordelingsplot:
Vælg Analyze ▶ Fit Y by X ▶ markér "Urenhed" og tryk på Y Response ▶ markér "Tilsætningsstof" og tryk på X Factor ▶ OK
Der fremkommer et "scatterplot", hvor man kan se afbildet de tre værdier af hver af de 4 tilsætningsstoffer
rød pil ved figur ▶ rullemenu Save ▶ Save Residual
Under data kommer nu en ekstra søjle med residualerne(overskrift "urenhed centered by tilsætningsstof").
Vælg Analyze ▶ Distribution ▶ Indsæt Residual søjlen i Y ▶ OK ▶ Cursor på rød pil ▶ "normal Quantile plot"

Der fremkommer et histogram med indtegnet normalfordelingskurve



Da vi jo kun har 12 residualer kan vi ikke forvente et særligt klart billede, men i princippet burde de på normalfordelings-plottet ligge nogenlunde på en ret linie. De ligger i hvert fald alle indenfor konfidens-grænserne, så vi kan tillade os at antage fordelingen er approksimativt normalfordelt.

Kontrol af varianshomogenitet:

X_i = mængden af uønsket stof ved tilsætning af stof T_i hvor $i \in \{1,2,3,4\}$

X_i antages approksimativt normalfordelt med middelværdien μ_i og spredning σ_i .

H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$ H: Mindst en varians er forskellig fra en af de øvrige

Sæt cursor på rød pil ved tegningen for Scatterplot, og vælg fra rullemenuen "UnEqual Variances".

Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	0,5146	3	8	0,6836
Brown-Forsythe	0,2404	3	8	0,8659
Levene	0,7464	3	8	0,5541
Bartlett	0,3036	3	.	0,8228

Warning: Small sample sizes. Use Caution.

Da P - værdierne alle er over 0.05 accepteres nulhypotesen H_0: De 4 varianser er ens.

b) H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ mod H: "mindst én middelværdierne er forskellig fra en af de øvrige".

Sæt cursor på rød pil ved scatterplot, og vælg fra rullemenuen "Mens/Anova".

Der fremkommer så en tegning og følgende udskrift:

Oneway Anova

Summary of Fit

Rsquare	0,64181
Adj Rsquare	0,507489
Root Mean Square Error	2,800298
Mean of Response	111,8
Observations (or Sum Wgts)	12

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
tilsætningsstof	3	112,40667	37,4689	4,7782	0,0342*
Error	8	62,73333	7,8417		
C. Total	11	175,14000			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
t1	3	110,000	1,6168	106,27	113,73
t2	3	108,000	1,6168	104,27	111,73
t3	3	113,167	1,6168	109,44	116,89
t4	3	116,033	1,6168	112,31	119,76

Std Error uses a pooled estimate of error variance

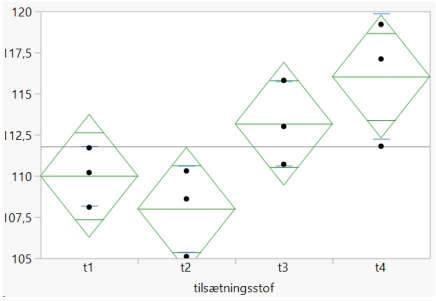
Af variansanalysetabellen fremgår, at P -værdi (Prob>F) = 0.0342 <0.05, dvs. nulhypotesen H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ forkastes (svagt)

Konklusion: De fire tilsætningsstoffer har ikke samme virkning.
Mindst 2 af middelurenhederne er forskellige.

Forklaring på de enkelte størrelser i variansanalysetabellen kan man finde i notatet "Supplement til Anvendt statistik" som findes på min hjemmeside.

c) Konfidensintervallerne ses i ovenstående udskrift, og umiddelbart ses, at der er et overlap mellem den mindste værdi t2 og t1 og t3.

For at få et lidt bedre skøn, ses på LSD-intervaller.



Tegningen der fremkom øverst på udskrift skal forstås på følgende måde:
"Diamanterne" angiver 95% konfidensintervaller.
Den midterste vandrette steg angiver gennemsnittet og de to andre vandrette streger angiver LSD -intervaller.
Af figuren ses derfor straks, ved at se på LSD intervallerne, at t2 er signifikant mindre end t4, mens det er vanskeligere at se om t2 og t3 kan adskilles.

LSD-intervaller fås ved fra rullemenuen at vælge
"Compare Means" ▶ Each Pair , Students t-tast
Blandt en række udskrifter findes denne

Connecting Letters Report

Level			Mean
t4	A		116,03333
t3	A	B	113,16667
t1		B	110,00000
t2		B	108,00000

Levels not connected by same letter are significantly different.

Konklusion: Man får den mindste urenhed, hvis man vælger enten T_2 eller T_1 (de kan ikke adskilles).

Af den forrige udskrift ses, at et estimat for mindsteværdi er 108.0
og et 95% konfidensinterval er [104.3 ; 111.7] ◆

13.3 FULDSTÆNDIGT RANDOMISERET BLOKFORSØG.

I forbindelse med planlægningen af et forsøg, kan man blive tvunget til at benytte forsøgsenheder, som er ret uensartede.
For at "dæmpe støjen" kan man inddele forsøgsenhederne i grupper (blokke), hvor de forsøgsenheder der ligger i samme blok er væsentlig mere ensartede end forsøgsenhederne i forskellige blokke. Man siger, at man har et fuldstændigt randomiseret blokforsøg, hvis hver behandling forekommer det samme antal gange (sædvanligvis netop én gang) i hver blok.
Eksempler på blokforsøg kan være

1) Af tidsmæssige grunde (forsøget skal udføres indenfor en given tidsramme) må man benytte 4 forskellige apparater. Påvirker apparaterne forsøgsresultaterne, kan det give en stor spredning ((stor "støj"). Dette kan bevirke, at man skal op på et urealistisk stort antal gentagelser for at kunne opnå den ønskede information.
   I stedet indfører man de 4 apparater som blokke.
2) Man er nødt til at benytte 5 forskellige råvareleverancer . Dette kan måske også give for stor spredning. Man indfører så de 5 råvareleverancer som 5 blokke.
3) I et landbrugsforsøg samles tilstødende arealer i blokke. Begrundelsen herfor er naturligvis, at arealer der ligger tæt ved hinanden er mere ensartede end arealer der ligger længere væk.

Eksempel 13.3 (randomiseret blokforsøg)
I nedenstående tabel er anført resultaterne af et fodringsforsøg med svin. Formålet med forsøget var at undersøge, hvorvidt en ændring af vitaminindholdet i foderet gav en forskel i svinenes vægtforøgelse. Vægtforøgelsen afhænger imidlertid også af det enkelte individs genetiske egenskaber. Et fuldstændigt randomiseret forsøg vil derfor sandsynligvis kunne bevirke, at forsøgsfejlens spredning bliver så stor, at intet kan påvises (forsøget drukner i støj). Da grise fra samme kuld må forventes at være mere ensartede, vælger man at lave et randomiseret blokforsøg med kuld som blokfaktor.
Lad der findes tre fodertyper A, B og C med forskelligt vitaminindhold.
Fra hvert af 4 forskellige kuld grise udtages nu 3 grise.
Et kuld vælges, og ved lodtrækning bestemmes hvilke af de 3 grise, der bliver fodret med fodertype A, hvilken med fodertype B og den sidste får naturligvis type C.
Et nyt kuld udtages, og man randomiserer igen foderet indenfor kuldet (blokken), osv.
Forsøgsresultaterne (vægtforøgelse i kg) var

		Fodertype		
		A	B	C
Kuld	1	7.0	14.0	8.5
	2	16.0	15.5	16.5
	3	10.5	15.0	9.5
	4	13.5	21.0	13.5

a) Test, om der er nogen væsentlig virkning af ændringen i foderets vitaminindhold.
b) Hvis der er en forskel, så skal man angive hvilken foderblanding, der giver den største vægtforøgelse.

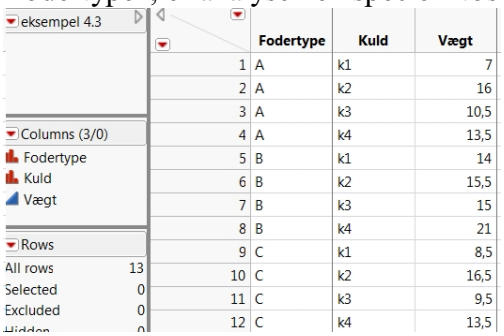
Løsning:

a) H_0: Foderblanding har ingen virkning på vægtforøgelsen

H: Foderblanding har virkning på vægtforøgelsen

Da vi jo har 2 faktorer, "kuld" og "fodertype", er analysen en speciel "tosidet variansanalyse".

Indtastning af data



	Fodertype	Kuld	Vægt
1	A	k1	7
2	A	k2	16
3	A	k3	10,5
4	A	k4	13,5
5	B	k1	14
6	B	k2	15,5
7	B	k3	15
8	B	k4	21
9	C	k1	8,5
10	C	k2	16,5
11	C	k3	9,5
12	C	k4	13,5

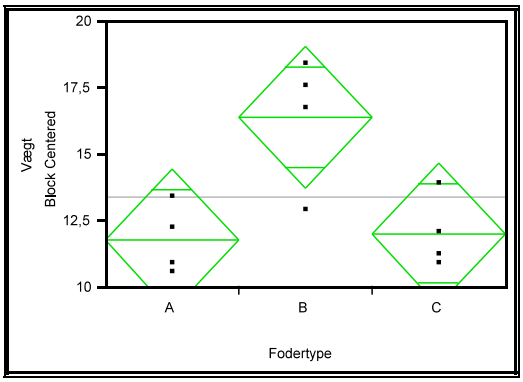
Vælg Analyze ▶ Fit Y by X ▶ markér "Vægt" og tryk på Y Response ▶ markér "Fodertype" og tryk på X Factor ▶ Marker "Kuld" og tryk på blok ▶ OK

Der fremkommer et "scatterplot", hvor man kan se afbildet de fire værdier for hver af de 3 fodertyper.

Rød pil ▶ "Means/Anova".

Der fremkommer så følgende tegning og udskrift:

Oneway Analysis of Vægt By Fodertype



Oneway Anova

Samarie of Fit

Rsquare	0,83413
Adj Rsquare	0,695904
Root Mean Square Error	2,168269
Mean of Response	13,375
Observations (or Sum Wgts)	12

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Fodertype	2	54,12500	27,0625	5,7563	0,0402
Kuld	3	87,72917	29,2431	6,2201	0,0285
Error	6	28,20833	4,7014		
C. Total	11	170,06250			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
A	4	11,7500	1,0841	9,097	14,403
B	4	16,3750	1,0841	13,722	19,028
C	4	12,0000	1,0841	9,347	14,653

Std Error uses a pooled estimate of error variance

Block Means

Kuld	Mean	Number
k1	9,8333	3
k2	16,0000	3
k3	11,6667	3
k4	16,0000	3

Da P - værdi for fodertype er 0.0402 < 0.05 forkastes nulhypotesen H_0: Ingen forskel på fodertyper, dvs. , at der på et signifikansniveau på 5 % er signifikant forskel på fodertyperne (mindst én afviger fra de øvrige).
Vi ser endvidere, at det var fornuftigt at dele op i kuld, da der også er signifikans for kuld.
Vi er imidlertid ikke interesseret i at finde ud af hvilket kuld der er det bedste, da vi jo blot har taget nogle tilfældige kuld ud. Konfidensintervallerne viser et svagt overlap.
Vælger nu LSD-intervaller
"Compare Means" ▶each pair students t-test

Comparisons for each pair using Student's t

Confidence Quantile

t	Alpha
2,44691	0,05

Connecting Letters Report

Level		Mean
B	A	16,375000
C	B	12,000000
A	B	11,750000

Levels not connected by same letter are significantly different.

Af LSD-intervallerne ses, at fodertype B giver den største vægtforøgelse.

Metode uden brug af færdigt "blokprogram"
Først foretages en ensidet variansanalyse med de tre fodertyper

Man noterer sig under faktor: SS = 54.125, df = 2 og MS= $s^2 = \frac{SS}{df} = 27.0625$

samt under error : SS_{error}=115.9375 med df = 9
Derefter en ensidet variansanalyse med de 4 kuld.
Man noterer sig SS_{blok} = 87.7291 med df = 3.

Det blokkene har forminsket støjen (error) findes ved ved SS = SSerror - SS_{blok} = 115.9375 - 87.7291 = 28.2084 med df = 9 - 3 = 6

Resultaterne indtastes i en såkaldt variansanalysetabel

Variation	SAK=SS	df	MS= $s^2 = \frac{SS}{df}$	F	P-værdi
Factor: foder	54.125	2	27.0625	$\frac{27.06}{4.70}$=5.756	$FCdf(5.756,\infty,2,6)$=0.0402
Block: Kuld	87.7291	3	29.2431	6.22	0.0284
Error	28.2084	6	4.70139		

H_0: Foder = 0 (Foder har ingen virkning) forkastes, da P-værdi = 0.0402 < 0.05
Konklusion: Der sker en væsentlig ændring i vægtforøgelsen ved at ændre foderblanding.
Da P-værdi for kuld er mindre end 5% har det haft betydning, at man har opdelt i blokke

13.4. Binomialfordelte variable

Ved analysen anvendes formlerne i afsnit 13.6.2.

Eksempel 13.4 (binomialfordelt variabel).

For hver af 6 leverancer af billige legetøjsbiler udtages en tilfældig prøve på 100 biler, og antallet af defekte biler taltes. Følgende resultater fandtes:

Leverance	1	2	3	4	5	6	7	8
Antal defekte biler	6	14	8	4	7	3	13	7

Foretag en statistisk analyse af, om procenten af defekte biler i de 8 leverancer kan antages at være den samme.

Løsning:

Lad X_i være antallet af defekte biler i leverance *i*.

Det antages, at X_i er binomialfordelt *b* (100, p_i).

H_0: $p_1 = p_2 = \ldots = p_8$

$$\hat{p}_1 = \frac{6}{100}, \hat{p}_2 = \frac{14}{100}, \ldots, \hat{p}_8 = \frac{7}{100}; \hat{p} = \frac{6+14+\ldots+7}{800} = \frac{64}{800} = 0.08$$

Da $n_i \cdot \hat{p} = 8 \in [5;95]$ er forudsætningen opfyldt (se evt. oversigten i afsnit 13.6.2)

$$\chi^2 = \frac{1}{0.08 \cdot (1-0.08)} 100\left((0.06-0.08)^2 + (0.14-0.08)^2 + \ldots + (0.07-0.08)^2\right) = \frac{1.08}{0.08 \cdot 0.92} = 14.67$$

χ^2 er χ^2 - fordelt med frihedsgradstallet $f = n - 1 = 7$

P - værdi $= P(\chi^2 > 14.67)$ 1 - ChiSquare Distribution 14,67, 7, $= 0.0405 =$

Da P-værdi < 0.05 forkastes nulhypotesen (svagt), dvs. vi har et (svagt) statistisk bevis for, at procenten af defekte biler i leverancen ikke er den samme.

13.5. Poissonfordelt variable

Ved analysen anvendes formlerne i oversigt 13.6.3.

Eksempel 13.5 (Poissonfordelt variabel)

Ved en optælling af hvide blodlegemer i en blodprøve med voluminet v fandtes for 6 *personer antallene* 14 , 28 ,18, 23,15 *og* 22.

Viser disse resultater, at den gennemsnitlige antal blodlegemer pr. volumenenhed er forskelligt for de 6 personer?

Løsning:

Lad X_1 vare antallet af hvide blodlegemer i en blodprøve for person 1

Lad X_2 vare antallet af hvide blodlegemer i en blodprøve for person 2

...

Lad X_6 vare antallet af hvide blodlegemer i en blodprøve for person 6.

X_i antages at være Poissonfordelt med middelværdi μ_i.

Begrundelse: Benyttes en kanyle til udtagning af blodprøven ankommer de hvide blodlegemer "tilfældigt" i tiden. Det mulige antal blodlegemer er næsten ubegrænset.

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$

Antal elementer i hver stikprøve er 1, dvs. i oversigten i afsnit 6.5.3 er $n_1 = n_2 = \ldots = n_6 = 1$ og $\bar{x}_1 = x_1 = 14, \bar{x}_2 = x_2 = 28, \ldots \bar{x}_6 = x_6 = 22$

Vi får $\bar{x} = \frac{14 + 28 + \ldots + 22}{6} = 20.0$.

Heraf ses, at $n_i \bar{x} \geq 5$, dvs. forudsætningen for at benytte oversigten i afsnit 6.6.3 er opfyldt.

$$\chi^2 = \frac{1 \cdot (14 - 20)^2 + 1 \cdot (28 - 20)^2 + \ldots + 1 \cdot (22 - 20)^2}{20} = 7.1$$

χ^2 er χ^2 - fordelt med frihedsgradstallet $f = n - 1 = 5$

Da P - værdi = $P(\chi^2 > 7.1)$ =1-ChiSquare Distribution (7,1,5) =0.2133 > 0.05 accepteres H_0 , det vil sige, at det ikke er påvist, at det gennemsnitlige antal hvide blodlegemer pr. volumenenhed er forskelligt for de 6 personer. ◆

13.6. OVERSIGT OVER CENTRALE FORMLER I KAPITEL 13

13.6.1 Oversigt over fremgangsmåde ved ensidet variansanalyse

Givet følgende skema:

	A_1	A_2		A_q
Observationer	x_{11} x_{21} . . . x_{n1}	x_{12} x_{22} . . . x_{n2}		x_{1q} x_{2q} . . . x_{nq}

Idet de q variables middelværdier kaldes $\mu_1, \mu_2, ..., \mu_q$ ønsker vi at teste nulhypotesen H_0: $\mu_1 = \mu_2 = ... = \mu_q$, mod H: "mindst én middelværdierne er forskellig fra en af de øvrige".

Data indtastes i 2 lister, med overskrift A og observationer.

Analyze ▶Fit Y by ▶X ▶marker observationer og tryk på Y Response ▶marker A og tryk på X Factor ▶OK

Der fremkommer bl.a. et scatterplot.

Sæt Cursor på rød pil ved scatterlot og vælg fra rullemenu Means/Anova

P-værdi aflæses. Hvis P-værdi < α forkastes H_0.

Konfidensintervaller

I Means/Anova ses konfidensintervaller

LSD - intervaller fås ved at vælge Compare Means

Er $|\bar{x}_1 - \bar{x}_2| > d_{LSD}$ hvor d_{LSD} er det dobbelte af radiua i LSD-intervallerne antages niveauerne A_1 og A_2 at være signifikant forskellige

(forudsat antallet af niveauer er lille (< ca 6).

Blokforsøg

		Faktor			
		A_1	A_2	. . .	A_q
Blokke	b_1	x_{11}	x_{12}		x_{1q}
	b_2	x_{21}	x_{22}		x_{2q}
	...	-	-		-
	b_n	x_{n1}	x_{n2}		x_{nq}

Indtaster og ordrer som før, men marker overskrigt "blokke" og tryk på blok

P-værdi aflæses. Hvis P-værdi < α forkastes H_0: $\mu_1 = \mu_2 = ... = \mu_n$

Konfidensintervaller

Samme ordrer som før

13.6.2 Oversigt over test af parametre $p_1, p_2, \ldots p_k$ for binomialfordelte variable.

$X_1, X_2 \ldots X_k$ er **binomialfordelt** henholdsvis $b(n_1, p_1),\ b(n_2, p_2), \ldots, b(n_k, p_k)$, hvor $n_1, n_2 \ldots n_k$ er kendte og $p_1, p_2 \ldots, p_k$ ukendte.
Observerede stikprøveværdier $x_1,\ x_2, \ldots, x_k$. Signifikansniveau er α.
Y er en statistisk variabel, der er χ^2 - fordelt med k - 1 frihedsgrader.

Forudsætning:	Nulhypotese	Beregning	P-værdi
Aproksimativ metode $n_1 \cdot \hat{p} \in [5; n_1 - 5]$ $n_2 \cdot \hat{p} \in [5; n_2 - 5]$ $\ldots \quad n_k \cdot \hat{p} \in [5; n_k - 5]$ hvor $\hat{p} = \frac{x_1 + x_2 + \ldots + x_k}{n_1 + n_2 + \ldots + n_k}$	$H_0: p_1 = p_2 = \ldots = p_k$	$\chi^2 = \frac{1}{\hat{p}(1-\hat{p})} \sum_{i=1}^{k} n_i (\hat{p}_i - \hat{p})^2$ hvor $\hat{p}_1 = \frac{x_1}{n_1}, \hat{p}_2 = \frac{x_2}{n_2}, \ldots, \hat{p}_k = \frac{x_k}{n_k}$	$P(Y > \chi^2)$

13.6.3. Oversigt over test af parametre $\mu_1, \mu_2, \ldots, \mu_k$ for Poissonfordelt variable.

$X_1, X_2 \ldots X_k$ er Poissonfordelt henholdsvis $p(\mu_1), p(\mu_2), \ldots, p(\mu_k)$,
hvor $\mu_1, \mu_2, \ldots, \mu_k$ er ukendte.
Signifikansniveau er α.
Der foreligger for hver af de variable X_i en stikprøve af størrelsen n_i med gennemsnit $\bar{x}_i$.
Y er en statistisk variabel, der er χ^2 - fordelt med k - 1 frihedsgrader..

Forudsætning	Nulhypotese	Beregning	H_0 forkastes
Approksimativ metode. $n_i \cdot \bar{x} \geq 5$	$H_0: \mu_1 = \mu_2 = \ldots = \mu_k$	$\chi^2 = \frac{\sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2}{\bar{x}}$ hvor $\bar{x} = \frac{n_1 \cdot \bar{x}_1 + n_2 \cdot \bar{x}_2 + \ldots + n_k \cdot \bar{x}_k}{n_1 + n_2 + \ldots + n_k}$.	P - værdi $< \alpha$ hvor P – værdi $= P(Y > \chi^2)$

OPGAVER

Opgave 13.1

Fire forskellige typer teknik til blanding af cement ønskes undersøgt med hensyn til resultatets trykstyrke. Følgende data blev opnået:

Blandingsteknik	Trykstyrke (psi)			
B1	3129	3000	2865	2890
B2	3200	3300	2975	3250
B3	2800	2900	2985	3050
B4	2600	2700	2600	2765

a) Kontroller om betingelserne om normalitet og varianshomogenitet er rimelig opfyldt.
b) Undersøg på et signifikansniveau på 5% om forskellen i blandingsteknik har betydning for trykstyrken.
c) Find i bekræftende fald den (de) blandingsteknik(er) der har størst trykstyrke, og angivet 95% konfidensinterval for middelværdien.

Opgave 13.2

I følgende tabel er angivet resultaterne af gentagne bestemmelser af blodets alkoholkoncentration (i promille) hos 6 forskellige personer efter indtagelsen af 6 cl. alkohol.

Person					
1	2	3	4	5	6
0.76	0.84	0.83	1.00	0.88	0.86
0.82	0.79	0.78	0.90	0.90	0.89
0.79	0.82	0.97	0.92	1.03	0.87
0.86	0.79	0.88	0.88	0.87	0.84

Analysens forudsætninger om varians og normalitet antages opfyldt.
Vurdér på grundlag af dette materiale en antagelse om, at alkoholkoncentrationen i blodet ikke afhænger af andre faktorer end den indtagne alkoho1mængde.

Opgave 13.3

Modstanden af 5 spoler måltes for at kontrollere, om spolerne har samme elektriske modstand. For hver spole måltes 4 uafhængige observationer:
Man fandt for hver spole følgende gennemsnit og spredning:

Spole nr.	Antal gentagelser	Gennemsnit	Spredning
1	4	15.2	0.1871
2	4	14.95	0.1225
3	4	14.8	0.1472
4	4	15,15	0.1826
5	4	14.7	0.2562

1) Undersøg om det kan antages, at de 5 spo1ers modstande er ens.
2) På alle 5 spoler er angivet, at modstanden er 15.0 Ohm. Undersøg under hensyntagen ti1 besvare1sen af spørgsmål 1) ved opstilling af et eller flere konfidensintervaller, om nog1e af spo1erne kan antages at have modstanden 15.0 Ohm og i bekræftende fald hvi1ke.

Opgave 13.4.
Følgende resultater blev opnået fra et eksperiment, hvor man ville undersøge om der var forskel på de resultater, som 5 analyseapparater gav, når man analyserede kvælstofindholdet i jordprøver.
På hver af 3 dage blev en portion jord udvalgt og delt i 5 dele, som ved lodtrækning blev givet til analyse i hver sin maskine. Resultaterne var:

		Maskiner				
		P	Q	R	T	U
	Tirsdag	376	379	399	373	376
	Onsdag	372	374	409	387	386
	Torsdag	332	339	365	350	342

Undersøg på dette grundlag om der er forskel mellem analyseapparaterne, og angiv i bekræftende fald hvilke der er forskellige.
Mener du, at det i denne situation var en god ide at foretage forsøget som et blokforsøg?

Opgave 13.5.
Fire forskellige produktionsmetoder P, Q, R, og T ønskes sammenlignet med hensyn til det procentiske udbytte ved udvinding af et metal fra et bestemt mineral. Da man ved forsøget er nødt til at benytte forskellige råvarepartierer, og er bange for, at det vil give stor spredning, vælger man at lave et fuldstændigt randomiseret blokforsøg med råvarepartier som blokke. Nedenstående skema angiver resultatet af dette forsøg.

	Metode P	Metode Q	Metode R	Metode T
Råvareparti1	2.5 2.7	4.7 3.3	2.8 3.3	5.5 5.0
Råvareparti 2	4.6 4.3	7.9 5.9	5.1 6.9	7.2 6.8
Råvareparti 3	4.7 3.9	4.7 4.4	4.4 3.7	6.4 5.7

Undersøg på grundlag af disse oplysninger, om der er forskel på metoderne.

Opgave 13.6
På en ingeniørskole ønsker man at sammenligne effektiviteten af undervisningen, når man underviser efter tre forskellige undervisningsmaterialer. En række studerende meldte sig frivilligt til forsøget. I det følgende er angivet 12 studerende ordnet efter studentereksamensgennemsnit.

Navn	JK	AL	TS	BS	DT	HN	MO	FD	PJ	KM	SR	RA
Snit	6.3	6.8	7.3	7.3	7.9	8.2	8.4	8.5	9.0	10.2	11.1	11.2

1) Hvordan ville du opdele disse studenter på tre hold med 4 på hver hold?
2) Hvordan ville du gøre det, hvis karaktererne gik fra 7.8 til 8.2 ?

Opgave 13.7.
I en virksomhed er på hvert af 3 skift arbejdsbetingelser og antal mennesker udsat for risiko tilsyneladende nogenlunde ens.
Ikke desto mindre synes følgende optælling at vise, at risikoen på skift 2 og 3 er større end på skift 1.

Skift	Antal arbejdsulykker
1	10
2	22
3	22

På grundlag af denne statistik finder man, at der bør gøres noget for at nedsætte risikoen i skift 2 og 3. Er dette statistisk velbegrundet.?

Opgave 13.8
5 typer vaccine mod en bestemt sygdom blev undersøgt ved, at 6 grupper på hver 200 forsøgsdyr (mus) blev udsat for smitte. De 5 af grupperne fik hver sin type vaccination, mens den sidste gruppe ikke blev vaccineret. Efter en passende tid undersøgte man hvor mange af de 200 dyr, der havde fået sygdommen. Følgende resultater fandtes:

Gruppe nr	1	2	3	4	5	6
Antal syge dyr	12	13	18	10	16	27

Vi ønsker at foretage en statistisk ana1yse af, om procenten af smittede dyr i de 6 grupper kan antages at være den samme.

14 2 FAKTORER PÅ 2 ELLER FLERE NIVEAUER TOSIDET VARIANSANALYSE

14.1 INDLEDNING.

Har man 2 kvalitative faktorer vil det også være naturligt at udføre en variansanlyse, men da man her kan risikere, at de to faktorer "spiller sammen" på en uventet måde, bliver forholdene noget mere kompliceret. Til gengæld kan begreberne her så umiddelbart generaliseres til forsøg med mere end 2 faktorer.

14.2 PLANLÆGNING AF FORSØG.

I dette afsnit benyttes følgende eksempel som illustration af begreberne.

Eksempel 14.1. 2 faktorer .
En bilfabrikant ønsker at finde ud af hvorledes 3 olieblandinger O_1, O_2, og O_3, og 2 karburatortyper K_1 og K_2 påvirker benzinforbruget.

◆

Vi har et forsøg med 2 **kvalitative faktorer:** olieblanding og karburator.
Faktoren "olieblanding" er på 3 **niveauer** O_1, O_2, og O_3, mens faktoren "karburator" har 2 niveauer nemlig K_1 og K_2 .
I alt er der altså 6 **behandlinger**.

14.2.1 En faktor ad gangen

I mange forsøgsvejledninger står, at man bør kun variere en faktor ad gangen. Alle andre faktorer end den udvalgte fastholdes på et bestemt niveau.
En forsøgsplan efter disse retningslinier kunne eksempelvis være som skitseret nedenfor, hvor hvert delforsøg er markeret med et ×, hvor man har valgt, at hvert niveau skal gentages mindst 4 gange.
Endvidere er det væsentligt, at hver af de indgående behandlinger har lige mange gentagelser.

		Karburator	
		K_1	K_2
Olieblanding	O_1	× × × ×	
	O_2	× × × ×	× × × ×
	O_3	× × × ×	

I dette eksempel, hvor der kun er 2 faktorer, vælger vi først at variere olieblandingen, mens den anden faktor fastholdes.
Idet vi har valgt først at fastholde karburatoren på niveauet K_1, kan forsøget udføres således:
12 af de 16 biler, som skal anvendes, udstyres med karburator K_1, og derefter (randomiseret) får 4 af disse biler olieblanding O_1, 4 andre biler olieblanding O_2, og de sidste 4 biler olieblanding O_3.

Efter at have kørt en udvalgt strækning måles benzinforbruget.
Derefter varieres den anden faktor (her karburator), mens olieblandingen fastholdes på O_2, dvs. de sidste 4 biler udstyres med karburator K_2 og olieblanding O_2.
Igen gennemkøres den udvalgte strækning, og benzinforbruget måles.
Indtegnes for hver karburator det gennemsnitlige benzinforbrug mod olie-blandingen kunne vi eksempelvis få tegningen på fig. 14.1.

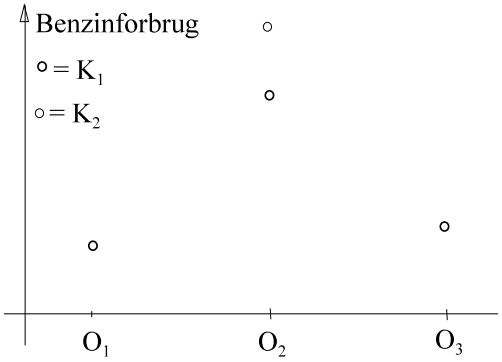


Fig 14.1 *Skitse af benzinforbrug*

Umiddelbart ses, at K_1 giver lavest benzinforbrug, og O_1 (eller O_3) skal foretrækkes.
Hvad nu med benzinforbruget i karburator K_2 , hvis vi anvender olieblanding O_1 eller O_3?
Kan man slutte, at benzinforbruget ved olieblanding O_1 og O_3 er lavere, når man bruger karburator K_1, end når man bruger karburator K_2?
Kun, hvis man ud fra tekniske eller andre grunde mener at vide, at "karburatorkurven" for K_2 er parallel med kurven for K_1, så er forsøgsplanen anvendelig, men ikke den bedste.

En statistisk set bedre forsøgsplan som endda ofte er mindre ressourcekrævende, er følgende:

14.2.2 Fuldstændig faktorstruktur

Denne plan består i, at hvert niveau af den ene faktor kombineres med ethvert niveau af den anden.
Planen kan skitseres således:

		Karburator	
		K_1	K_2
Olieblanding	O_1	× ×	× ×
	O_2	× ×	× ×
	O_3	× ×	× ×

Her er hver af de 6 behandlinger gentaget 2 gange, dvs. i alt er der udført 12 delforsøg.
Hermed er kravet opfyldt om at hvert niveau skal gentages mindst 4 gange.

Det er vigtigt, at hver behandling har lige mange gentagelser
I " en faktor ad gangen" var vi tvunget til at udføre 16 delforsøg, mens vi kun skal lave 12 delforsøg i det "fuldstændige faktorforsøg".
Vi kan altså nøjes med færre delforsøg, når vi laver et fuldstændigt faktorforsøg.

Indtegnes for hver karburator det gennemsnitlige benzinforbrug mod olie-blandingen, kan det eksempelvis vise sig, at man får figur 14.2.

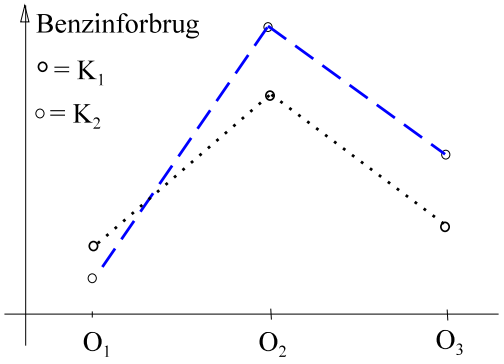


Fig 14.2. *Vekselvirkning*

Vi ser, i modstrid med hvad vi antog ud fra "en faktor ad gangen forsøget", at kombinationen af katalysator K_2 og olieblanding O_1 giver det laveste benzinforbrug.

Det ses, at de to kurver **ikke er parallelle**. Dette kunne være tilfældigt og blot skyldes forsøgets "støj", men det kunne også være signifikant, og derfor være udtryk for en såkaldt **"vekselvirkning"**.
En model uden vekselvirkning (kurverne tilnærmelsesvis parallelle) siges at være **additiv.**

14.3. FORMLER OG METODE

Vi vil i det følgende kun analysere forsøg med en fuldstændig faktorstruktur, og hvor hver behandling har lige mange gentagelser (samme antal delforsøg i hver "celle").
Et eksempel på et sådant forsøg er

Eksempel 14.2. 2 faktorer (fortsat fra eksempel 14.1) .
En bilfabrikant ønsker at finde ud af, hvorledes 3 olieblandinger O_1, O_2, og O_3, og 2 karburatortyper K_1 og K_2 påvirker benzinforbruget.
Forsøgsresultaterne er følgende:

		Karburator			
		K_1		K_2	
Olieblanding	O_1	830	860	810	840
	O_2	940	990	1050	1020
	O_3	855	815	930	910

Angiv hvilke kombinationer af karburator og olieblanding der giver det laveste forbrug, og giv et estimat for dette forbrug.
Symbolik:
Lad os kalde rækkefaktoren for R , antal rækkeniveauer r , søjlefaktoren for C, antal søjleniveauer for q og antal gentagelser af hver behandling n. Det totale antal delforsøg er følgelig $N = r \cdot q \cdot n$.
I eksempel 14.4 er R = "olieblanding", $r = 3$, C = " karburator" , $q = 2$, $n = 2$ og $N = 12$.

Forudsætninger
Disse er de samme som ved den ensidede variansanalyse. Analysen er også her robust overfor afvigelser fra normalitet og varianshomogenitet, blot antallet af gentagelser i hver celle er den samme.

Skitse af fremgangsmåde ved testning
Først testes om modellen er additiv, dvs. om den er uden en signifikant vekselvirkning.
Nulhypotesen skrives kort $H_0: R \cdot C = 0$ (faktorerne vekselvirker ikke)
og den alternative hypotese $H: R \cdot C \neq 0$ (faktorerne vekselvirker)
For at kunne beregne en P-værdi beregnes først et estimat for den fælles (poolede) varians s_0^2 på samme måde som det skete i forrige kapitel.
Derefter beregnes et estimat for en varians s_{RC}^2 som vil være lig s_0^2 hvis H_0 er sand, men væsentlig større hvis H er sand

P-værdien beregnes som en F-test ud fra brøken $F = \frac{s_{RC}^2}{s_0^2}$.

1) Er P-værdi < α (signifikansniveau) forkastes H_0 dvs. faktorerne vekselvirker.
   Ved hjælp af konfidensintervaller for alle $r \cdot q$ celler søger man at finde den optimale kombination af faktorer.

2) Er P-værdi > α accepteres H_0 dvs. faktorerne vekselvirker ikke (model er additiv).
   Man tester nu nulhypoteserne
   $H_0: R = 0$ (rækkefaktor har ikke en virkning)
   $H_0: C = 0$ (søjlefaktor har ikke en virkning)
   Da man må antage, at de to varianser s_{RC}^2 og s_0^2 er nogenlunde ens, pooles de sammen til et nyt (bedre) estimat s_m^2 for forsøgsfejlen (error).
   Variansen s_m^2 benyttes nu ved beregning af P-værdierne for de to hypoteser.
   a) Finder man at såvel $H_0: R = 0$ forkastes, som $H_0: C = 0$ forkastes vil man sædvanligvis opstille konfidensintervaller til bestemmelse af de to faktorers optimale niveau.
      Hertil benyttes s_m^2, da det jo giver et sikrere konfidensinterval.
   b) Finder man at eksempelvis $H_0: R = 0$ forkastes, mens $H_0: C = 0$ accepteres pooles varianserne s_{RC}^2, s_0^2 og s_c^2 sammen til en fælles varians, og denne benyttes ved opstillingen af konfidensintervaller for R.

Beregninger: Hvorledes man foretager testen ved hjælp af en lommeregner der kun kan beregne gennemsnit og spredning er beskrevet i notatet "Supplement til Anvendt statistik" som findes på min hjemmeside. Da specielt beregningerne af SS=SAK'erne er temmelig omfattende, og næppe giver en dybere forståelse ,vil vi dog sædvanligvis benytte programmer som findes i SAS.JMP hertil. Programmerne dækker dog ikke alle muligheder, så i enkelte tilfælde må man derfor benytte formlerne for eksempelvis konfidensintervaller.

Anskuelig forklaring på hvorledes man kan beregne vekselvirkning: Her gives kun en kort forklaring, som kan tjene til at forstå baggrunden for beregningerne, der i øvrigt med fordel kan foretages af et statistikprogram .
I nedenstående skema er skitseret et forsøg med 2 faktorer R og C. R er på 3 niveauer, og C er på 4 niveauer. Der er 2 gentagelser af hver "behandling"(treatment).

	C_1	C_2	C_3	C_4
R_1	1 3	2 6	1 5	6 8
R_2	7 11	8 14	8 12	13 15
R_3	2 6	5 7	3 7	8 10

For hver af de 12 celler kan man udregne et skøn for spredningen. Hvis man forudsætter at spredningen er nogenlunde den samme i alle 12 tilfælde, kan man poole de 12 s^2 sammen til et fælles skøn s_0 for spredningen på forsøgsfejlen (støjen). Den vil have 12 frihedsgrader, da hvert enkelt s har 1 frihedsgrad.
I nedenstående skema er beregnet gennemsnit for hver celle, hver række, hver søjle og totalt.

	C_1	C_2	C_3	C_4	Gennemsnit
R_1	2	4	3	7	4
R_2	9	11	10	14	11
R_3	4	6	5	9	6
Gennemsnit	5	7	6	10	7

Tallene er konstrueret således, at vi har en helt præcis model uden vekselvirkning ($R_2 = R_1 + 7$, $R_3 = R_1 + 2$). For en sådan model gælder helt præcist, at resultatet i celle (i, j) fås af formlen $RC_{i,j}=R_i + C_j$ - totale gennemsnit.
Eksempel: $RC_{2,3} = 10$ og $R_2 + C_3$ - totale gennemsnit = 11 + 6 - 7 = 10.
I praksis vil dette naturligvis aldrig være tilfældet på grund af den tilfældige variation (støj), men udregnes kvadratet på afvigelserne (SAK), og disse afvigelser ikke er større end hvad er rimeligt i forhold til støjen (s_0), vil vi kunne konkludere at der ikke kan konstateres nogen vekselvirkning.

14.4 BEREGNING AF TOSIDET VARIANSANALYSE

Som nævnt er der 2 hovedtilfælde, nemlig om der konstateres vekselvirkning eller ej.
Er modellen additiv er der igen 2 tilfælde, nemlig om begge faktorer har en signifikant virkning eller kun den ene faktor har en signifikant virkning.
Vi ser her bort fra den situation, at ingen faktorer har en signifikant virkning. Det kan naturligvis sagtens forekomme, men den situation kræver ingen nærmere forklaring.

I afsnit 14.7 er givet en oversigt over hvorledes beregningerne skal udføres med SAS.JMP.

I det følgende anvendes denne oversigt til at regne 3 eksempler svarende til hver af de tre situationer.

Eksempel 14.2 (fortsat) **Model med vekselvirkning.**
En bilfabrikant ønsker at finde ud af, hvorledes 3 olieblandinger O_1, O_2, og O_3, og 2 karburatortyper K_1 og K_2 påvirker benzinforbruget.
Forsøgsresultaterne er følgende:

		Karburator	
		K_1	K_2
Olieblanding	O_1	830 860	810 840
	O_2	940 990	1050 1020
	O_3	855 815	930 910

a) Kontroller om betingelserne for normalitet og varianshomogenitet er rimelig opfyldt
b) Angiv hvilke kombinationer af karburator og olieblanding der giver det laveste forbrug, og giv et estimat og et konfidensinterval for dette forbrug.

Løsning:

Data indtastes på sædvanlig måde .

Karburator	olieblanding	benzinforbrug	
k1	o1	830	
k1	o1	860	
k1	o2	940	
k1	o2	990	
k1	o3	855	
k1	o3	815	
k2	o1	810	osv.

a) **Kontrol af normalfordeling**

For at kunne beregne residualer fra hver celle dannes en ekstra søjle, som angiver cellenummer.

Karburator	Olieblanding	Benzinforbrug	celler
k1	o1	830	c1
k1	o1	860	c1
k1	o2	940	c2
k1	o2	990	c2
k1	o3	855	c3
k1	o3	815	c3
k2	o1	810	c4
k2	o1	840	c4
k2	o2	1050	c5
k2	o2	1020	c5
k2	o3	930	c6
k2	o3	910	c6

Vælg Analyze ▶ Fit Y by X ▶ markér "Benzinforbrug" og tryk på Y Response ▶ markér "celler" og tryk på X Factor ▶ OK

Der fremkommer et "scatterplot"

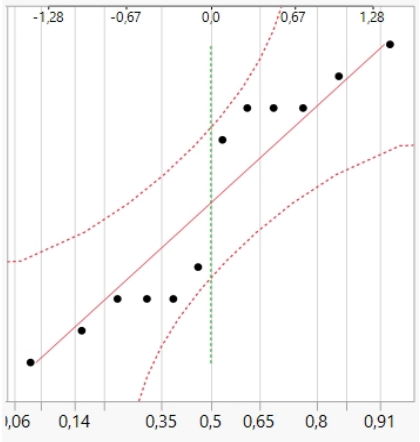
rød pil ved figur ▶ rullemenu Save ▶ Save Residual

Under data kommer nu en ekstra søjle med residualerne(overskrift "urenhed centered by tilsætningsstof").

Vælg Analyze ▶ Distribution ▶ Indsæt Residual søjlen i Y ▶ OK ▶ Cursor på rød pil ▶ "normal Quantile plot"

Der fremkommer et histogram med indtegnet normalfordelingskurve

Da vi jo kun har 12 residualer kan vi ikke forvente et særligt klart billede, men i princippet burde de på normalfordelings-plottet ligge nogenlunde på en ret linie. De ligger i hvert fald alle indenfor konfidens-grænserne, så vi kan tillade os at antage fordelingen er approksimativt normalfordelt.



Varianshomogenitet.

Lad varianserne i de 6 celler være σ_1,σ_2, σ_3, σ_4, σ_5 og σ_6.

H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2 = \sigma_6^2$ H: Mindst en varians er forskellig fra en af de øvrige

Da varianserne i hver af 6 celler skal være ens, karakteriserer man disse ved at vælge "celler"

Vælg Analyze ▶ Fit Y by X ▶ markér "benzinforbrug" og tryk på Y Response ▶ markér "celler" og tryk på X Factor ▶ OK cursor på rød pil på tegning og vælg "Unequal Variances"

Oneway Analysis of Benzinforbrug By Celler

Tests that the Variances are Equal

Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	0,0000	-1	0	0,0000*
Brown-Forsythe	.	5	6	.
Levene	.	5	6	.
Bartlett	0,1354	5	.	0,9842

Warning: Small sample sizes. Use Caution.

Da Bartletts test giver en P-værdi= 0.9642 > 0.05 fås en accept af nulhypotesen.
Kravet er rimeligt opfyldt.

b) Først testes $H_0: R*C=0$ mod $H: R*C \neq 0$

Vælg Analyze▶Fit Model ▶Indsæt for Y: benzinforbrug ,Indsæt for ADD: karburator og olieblanding▶Indsæt for CROSS: karburator, olieblanding (marker begge) ▶Emphasis's rullemenu: vælg Minimal Report▶Run

(Minimal report er kun valgt for at undgå nogle i denne forbindelse overflødige figurer)

Man får (blandt andet)

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
karburator	1	1	6075,000	10,1250	0,0190
Olieblanding	2	2	58716,667	48,9306	0,0002
karburator*Olieblanding	2	2	6450,000	5,3750	0,0460

Ud for "Karburator*olieblanding" findes P - værdi = 0.0460.
Da P - værdi = 0.0460 < 0.05 forkastes H_0 (svagt).

Konklusion: Begge faktorer har en virkning i form af en vekselvirkning.
For at finde hvilke kombinationer der giver lavest benzinforbrug vælges

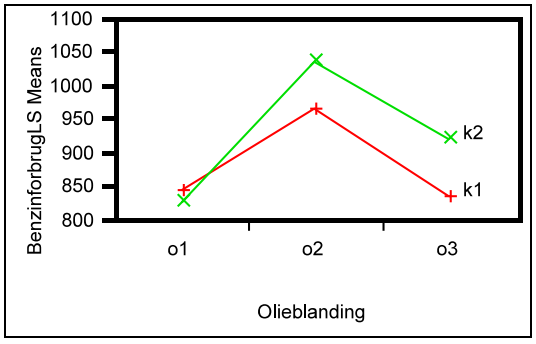
"Effect Details" ▶ rød pil ved "karburator x olieblanding ▶ Vælg "LS means Plot"

Vi får følgende Udskrift+tegning :

Karburator*Olieblanding
Least Squares Means Table

Level	Least Sq Mean	Std Error
k1,o1	845,0000	17,320508
k1,o2	965,0000	17,320508
k1,o3	835,0000	17,320508
k2,o1	825,0000	17,320508
k2,o2	1035,0000	17,320508
k2,o3	920,0000	17,320508

LS Means Plot



Umiddelbart ses af figuren, at man ikke bør vælge olieblanding O_2.
Derimod er det uklart hvilken af kombinationer (se tabellen) med de mindste "means", der giver det laveste olieforbrug.
Dette kan afklares ved på ovennævnte rullemenu at vælge "LSMeans students t"
Det giver en stor tabel (som kan fjernes ved med cursor på overskrift, højre musetast at fjerne markeringen ved "Crostab Report).
Under den findes følgende lille tabel

Level			Least Sq Mean
k2,o2	A		1035,0000
k1,o2		B	965,0000
k2,o3		B	920,0000
k1,o1		C	845,0000
k1,o3		C	835,0000
k2,o1		C	825,0000

Levels not connected by same letter are significantly different

Heraf ses, at kombinationen K_2 O_1 giver det laveste benzinforbrug (825), men, at der ingen signifikant forskel er mellem K_2 O_1 , K_1 O_3 og K_1 O_1.

Konfidensintervaller

Ønskes fundet 95% konfidensintervaller
musetast i skemaet ved Karburator*Olieblanding ▶tryk højre musetast ▶ Vælg "Columns" ▶vælg "lower" og derefter "upper".

Karburator*Olieblanding

Least Squares Means Table

Level	Least Sq Mean	Std Error	Lower 95%	Upper 95%
k1,o1	845,0000	17,320508	802,61824	887,3818
k1,o2	965,0000	17,320508	922,61824	1007,3818
k1,o3	835,0000	17,320508	792,61824	877,3818
k2,o1	825,0000	17,320508	782,61824	867,3818
k2,o2	1035,0000	17,320508	992,61824	1077,3818
k2,o3	920,0000	17,320508	877,61824	962,3818

Heraf ses, at for kombinationen K_2 O_1 er konfidensintervallet = [782.6 ; 867.4]

◆

Eksempel 14.3 Additiv model: To signifikante hovedvirkninger

I forbindelse med nogle brudstyrkebestemmelser for Portland-cement udføres et fuldstændigt randomiseret forsøg til undersøgelse af middelbrudstyrkens afhængighed af cementblandere og cementknusere.
Med hver af 3 cementblandere udstøbtes efter blanding med vand 12 cementterninger, som efter en uges lagring underkastedes en brudstyrkeprøve ved hjælp af en af 4 cementknusere.
Forsøgsresultaterne var:

		Cementknusere											
		1			2			3			4		
Cementblandere	1	147	175	130	99	85	75	67	23	35	215	97	180
	2	221	155	173	141	110	155	85	55	81	161	167	177
	3	123	85	153	137	143	82	67	25	83	135	91	129

Forudsætningerne for en variansanalyse antages opfyldt.
Angiv hvilke kombinationer af cementblander og cementknuser, der giver den største brudstyrke, og giv et estimat og et 95% konfidensinterval for denne største middelbrudstyrke.

Løsning:

Lad starten af indtastningen i regnearket være

cementblandere	cementknusere	brudstyrke
c1	k1	147
c1	k1	175
c1	k1	130
c1	k2	99
c1	k2	85

osv.

1) Variansanalysetabel opstilles.

Vælg Analyze▶Fit Model ▶Indsæt for Y: Brudstyrke ,Indsæt for AD:Cementblandere og Cementknusere▶Indsæt for CROSS: Cementblandere,Cementknusere (marker begge) ▶Run

Der fremkommer en række figurer og tabeller.

Blandt disse er følgende tabel

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Cementblander	2	2	8706,056	4,9116	0,0163*
Cementknuser	3	3	51995,222	19,5557	<,0001*
Cementblander*Cementknuser	6	6	7122,611	1,3394	0,2787

a) $H_0: R*C = 0$ (Ingen signifikant vekselvirkning) H: $R*C \neq 0$

For "Cementblandere*cementknusere" findes P - værdi = 0.2787.

Da P - værdi = 0.2787 > 0.05 accepteres H_0.

Konklusion: Vi antager i det følgende, at vekselvirkningen er forsvindende.

b) Vekselsvirkningen "pooles" ned i "error"".

Marker under den øverste Source: Cementblander*Cementknusere▶Remove

Blandt mange tabeller findes

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Cementblander	2	2	8706,056	4,5994	0,0181*
Cementknuser	3	3	51995,222	18,3125	<,0001*

H_0: C = 0 (Cementknuser har ingen virkning) forkastes, da P-værdi = 0.000 < 0.05

H_0: R = 0 (Cementblander har ingen virkning) forkastes, da P-værdi = 0.0181 < 0.05

Konklusion: Cementknuserne har en stærk signifikant virkning,
Cementblanderne har en signifikant virkning,

2) Under "cementblandere" kan på samme måde som i forrige eksempel bl.a. findes følgende tabeller

Level		Least Sq Mean
c2	A	140,08333
c1	B	110,66667
c3	B	104,41667

Levels not connected by same letter are significantly different.

Heraf ses, at cementblander 2 må foretrækkes

Under cementknusere fås

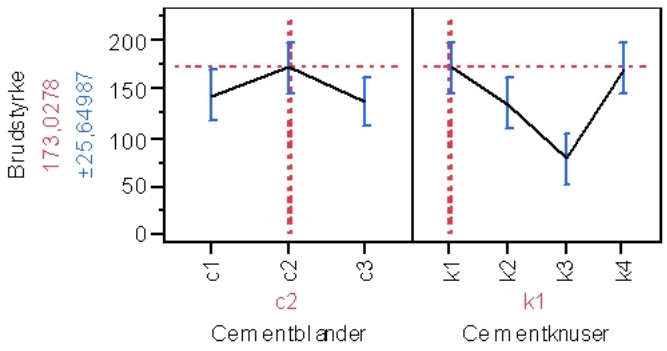
Level			Least Sq Mean
k1	A		151,33333
k4	A		150,22222
k2		B	114,11111
k3		C	57,88889

Levels not connected by same letter are significantly different
Cementknuser 1 og 4 må foretrækkes.
Konklusion: Størst middelbrudstyrke fås i kombinationen
cementknuser 1 og cementblander 2 eller cementknuser 4 og cementblander 2

Et estimat $\widetilde{\mu}_{12}$ for største middelbrudstyrke på basis af cementknuser 1 og cementblander 2:

Rød pil ved Response Brudstyrke ▶ Factor profiling ▶ Profiler

Ved de fremkomne figurer flyttes linier



$\widetilde{\mu}_{12} = 173.0$

95% Konfidensinterval : $[173.03 - 25.65; 173.03 + 25.65] = [147.4; 198.7]$

Eksempel 14.4 Additiv model: Een signifikant hovedvirkning

Samme problem som i eksempel 14.3, blot er tallene for cementblander 2 formindsket med 10.
Forsøgsresultaterne var:

		Cementknusere											
		1			2			3			4		
Cementblandere	1	147	175	130	99	85	75	67	23	35	215	97	180
	2	211	145	163	131	100	145	75	45	71	151	157	167
	3	123	85	153	137	143	82	67	25	83	135	91	129

Forudsætningerne for en variansanalyse antages opfyldt.

Angiv hvilke kombinationer af cementblander og cementknuser, der giver den største brudstyrke, og giv et estimat og et 95% konfidensinterval for denne største middelbrudstyrke.

Løsning:

a) $H_0 : R * C = 0$ (Ingen signifikant vekselvirkning) H: R*C=0

Fremgangsmåden er ganske analoge til eksempel 14.3.

For "interaction" findes P - værdi = 0.2787.

Da P - værdi = 0.2787 > 0.05 accepteres H_0.

Konklusion: Vi antager i det følgende, at vekselvirkningen er forsvindende.

b) Marker under Source: Cementblander*Cementknusere ▶ Remove
Får nu

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Cementblander	2	2	4299,389	2,2713	0,1206
Cementknuser	3	3	51995,222	18,3125	<,0001*

H_0: C = 0 (Cementknuser har ingen virkning) mod H: C $\neq$ 0
Da P-værdi for C er under 0.0001 må H_0: C = 0 forkastes, dvs. Cementknusere har en (stærk) virkning
H_0: R = 0 (Cementblander har ingen virkning) mod H: R $\neq$ 0
H_0: R = 0 (Cementblander har ingen virkning) accepteres, da P-værdi = 0.1206 > 0.05
Konklusion: Cementknuserne har en (stærk) virkning
Cementblandere har ingen virkning

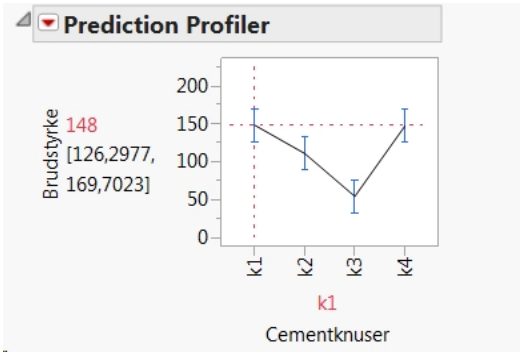
Marker under Source: Cementblander ▶ Remove
Under "effect details" ▶ rød pil ved "Cementknuser" ▶ vælges lsMean Students t-test
Det resulterer i følgende tabel:

Level		Least Sq Mean
k1	A	148,00000
k4	A	146,88889
k2	B	110,77778
k3	C	54,55556

Levels not connected by same letter are significantly different.

Konklusion: Cementknuser 1 og 4 må foretrækkes

Konfidensinterval:
Rød pil ved Response Brudstyrke ▶ Factor profiling ▶ Profiler
Ved den fremkomne figur flyttes linie.



Et estimat for største middelbrudstyrke: 148 95% konfidensinterval [126.3 ; 169.7]

14.5 FULDSTÆNDIGT RANDOMISERET BLOKFORSØG.

Som beskrevet i kapitel 4.3, kan man blive tvunget til at benytte forsøgsenheder, som er ret uensartede. Derved får den tilfældige forsøgsfejl en relativ stor spredning (stor "støj").
For at "dæmpe støjen" kan man inddele forsøgsenhederne i grupper (blokke), hvor de forsøgsenheder der ligger i samme blok er væsentlig mere ensartede end forsøgsenhederne i forskellige blokke.
Til illustration heraf, så betragter vi igen forsøget beskrevet i eksempel 14.1.

Eksempel 14.5 (randomiseret blokforsøg). En bilfabrikant ønsker at finde ud af, hvorledes 3 olieblandinger O_1, O_2, og O_3, og 2 karburatortyper K_1 og K_2 påvirker benzinforbruget. Forsøget planlægges som et fuldstændigt faktorforsøg idet hvert niveau skal gentages mindst 4 gange. Dette betyder at der skal udføres 12 delforsøg.
Et delforsøg med én bil tager 1 dag.(1 tank = 40 liter: Kører ca. 15 km/l så 40 liter = 600 km, hvilket giver ca. 7 timer med 80 km/time). Af tidsmæssige grunde kan man ikke benytte 12 dage til forsøget. Der benyttes 2 biler med tilhørende chauffør, hvilket forkorter forsøgstiden til 6 dage. Da de to biler (med tilhørende chauffør) kan frygtes at give systematisk forskellige resultater, ønskes foretaget et **randomiseret blokforsøg** med biler som blokke.

1) Beskriv hvorledes en randomisering kunne tænkes at foregå.
2) Opdelingen i 2 blokke resulterede i følgende plan med tilhørende benzinforbrug

bil 1	dag 1 O_1K_2 810	dag 2 O_3K_2 910	dag 3 O_2K_2 1020	dag 4 O_1K_1 830	dag 5 O_2K_1 940	dag 6 O_3K_1 815
bil 2	dag 1 O_1K_1 860	dag 2 O_3K_2 930	dag 3 O_1K_2 840	dag 4 O_2K_1 990	dag 5 O_3K_1 855	dag 6 O_2K_2 1050

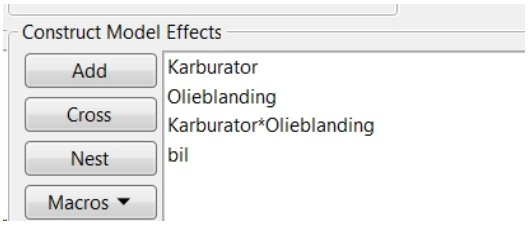
Beregn om man på dette grundlag kan vise, at der er vekselvirkning mellem olieblanding og karburator.

Løsning:

1) Randomisering: To dåser mærkes henholdsvis bil1 og bil2. Behandlingen O_1K_1 skrives på 2 sedler som anbringes i hver sin dåse, behandlingen O_1K_2 skrives på 2 sedler som anbringes i hver sin dåse osv.
Man trækker nu først de 6 sedler fra dåse med mærket bil 1.
Lad den første seddel der trækkes være O_1K_2. Det betyder nu, at bil 1 skal forsynes med karburator 2 og olieblanding 1 og køre dag 1. Lad den næste seddel der trækkes være O_3K_2 . Det betyder tilsvarende at bil 1 skal forsynes med karburator 2 og olieblanding 3 og køre dag 2. Således fortsættes indtil alle 6 sedler er udtrukket
Derefter fortsættes med at trække sedler fra dåsen med bil 2.
2) Data indtastes

	Karburator	Olieblanding	Benzinforbrug	bil
1	k1	o1	830	bil1
2	k1	o1	860	bil2
3	k1	o2	940	bil1
4	k1	o2	990	bil2
5	k1	o3	855	bil2
6	k1	o3	815	bil1
7	k2	o1	810	bil1
8	k2	o1	840	bil2
9	k2	o2	1050	bil2
10	k2	o2	1020	bil1
11	k2	o3	930	bil2
12	k2	o3	910	bil1

Vælg Analyze▶Fit Model ▶Indsæt for Y: Benzinforbrug ,Indsæt for AD:Cementblandere og Cementknusere▶Indsæt for CROSS: Cementblandere,Cementknusere (marker begge) ▶Indsæt for AD: bil ▶Run



Bemærk: Vi antager altid, at **blokke ikke vekselvirker med faktorerne**, idet vi forudsætter, at den ene blok (eksempelvis bil 1) bidrager med en systematisk højere resultat end den anden blok (eksempelvis at bil 1 på alle dage giver et større benzinforbrug end bil 2).
Der fremkommer bl.a. følgende udskrift

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Karburator	1	1	6075,000	113,9062	0,0001*
Olieblanding	2	2	58716,667	550,4687	<,0001*
Karburator*Olieblanding	2	2	6450,000	60,4687	0,0003*
bil	1	1	3333,333	62,5000	0,0005*

Idet vi som sædvanlig antager at signifikansniveauet er 5 % fås, at da P - værdi = 0.03% <5% forkastes H_0

Konklusion: Begge faktorer har en virkning i form af en vekselvirkning.

P-værdien er væsentlig mindre end i eksempel 14.2, dvs. blokken (bilerne) har haft betydning

Bemærk: Selv om analysen viser, at blokkene mod forventning ikke kan antages at have betydning, må man **ikke poole blokkene ned**, da det svarer til, at man analyserer forsøget som om det var et fuldstændigt randomiseret forsøg.

◆

14.6 Andre varianter af variansanalyse

14.6.1 Manglende observationer (ubalancerede forsøg).

Vi har hidtil forudsat, at forsøgsplanen var et fuldstændigt faktorforsøg med lige mange gentageser af hver behandling. Dette vil i praksis ikke altid være tilfældet. Mislykkes således et delforsøg, opdager man det måske først så sent, at det ikke er muligt at foretage et nyt forsøg under de samme betingelser. Man har så et ikke-balanceret forsøg. Analysen foretages på samme måde, som hvis dataene var balancerede, men forsøget er nu mere følsomt overfor, om forudsætningen om varianshomogenitet er opfyldt.

14.6.2 Trinvis variansanalyse

Trinvise niveauer. At en faktor B er undertrin (engelsk: nested factor) til faktor A's niveauer, betyder at ethvert niveau af B kun kombineres med ét niveau af A.

Et eksempel på en trinvis variansanalyse er følgende

Eksempel 14.6 (trinvis variansanalyse)

I hvert af 5 tilfældigt valgte autoriserede reparationsværksteder udvalgtes tilfældigt 2 mekanikere, og for hver af disse måltes 3 gange den tid, som en bestemt reparation tog. Følgende resultater fandtes (kodede tal):

Værksteder							
1		2		3		4	
Mekaniker		Mekaniker		Mekaniker		Mekaniker	
1	2	3	4	5	6	7	8
6 2 0	1 10 0	10 7 12	4 1 9	10 11 6	7 12 15	5 10 8	0 8 6

Udfør en trinvis "random factor" variansanalyse og drag konklusioner.

Løsning

Data indtastes

værksted	mekaniker	tid
v1	m1	6
v1	m1	2
v1	m1	0
v1	m2	1
v1	m2	10
v1	m2	0
v2	m3	10
v2	m3	7
v2	m3	12
v2	m4	4
v2	m4	1
v2	m4	9
v3	m5	10
v3	m5	11
v3	m5	6
v3	m6	7
v3	m6	12
v3	m6	15
v4	m7	5
v4	m7	10
v4	m7	8
v4	m8	0
v4	m8	8
v4	m8	6

Vælg Analyze ▶ Fit Model ▶ Indsæt for Y: tid ,Indsæt for ADD: værksted og mekaniker ▶tryk på mekaniker under effects og på værksted under "select columns ▶ Tryk på Nest ▶ Emphasis's rullemenu: vælg Minimal Report ▶Run

Resultat:

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
værksted	3	3	150,00000	3,6585	0,0351*
mekaniker[værksted]	4	4	60,66667	1,1098	0,3860

Det ses, at der er forskel på værkstederne, da P-værdi = 0.035 < 0.05.

Derimod er der ikke forskel på mekanikerne.

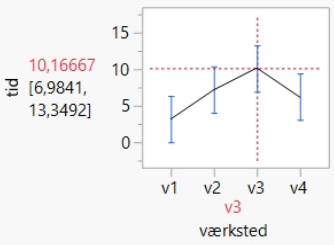
Pooler man mekanikerne væk, fås ved students t-test osv

Level		Least Sq Mean
v3	A	10,166667
v2	A B	7,166667
v4	A B	6,166667
v1	B	3,166667

Levels not connected by same letter are significantly different.

Heraf ses, at v3,v2 og v4 ikke kan adskilles. v1 er det dårligste værksted.

Prediction Profiler



I visse situationer, kan man som det følgende eksempel viser frygte, at der er mere end 1 betydende støjkilde. Et såkaldt romersk kvadratforsøg kan så være en mulighed for at dæmpe støjen.

Eksempel 14.7 (romersk kvadratforsøg) En bilfabrikant ønsker at finde ud af, hvorledes 3 olieblandinger O_1, O_2, og O_3, og 2 karburatortyper K_1 og K_2 påvirker benzinforbruget. Forsøget planlægges som et fuldstændigt faktorforsøg idet hvert niveau skal gentages mindst 4 gange. Dette betyder at der skal udføres 12 delforsøg.
Hvert delforsøg med én bil tager 1 dag. Man frygter, at forskellige dage med deres forskellige vejrforhold kan give stor variation i resultaterne. Endvidere frygter man også, at forskellige biler (med tilhørende chauffør) frygtes give systematisk forskellige resultater.
Da der er i alt 6 forskellige behandlinger ønskes foretaget et **romersk kvadratforsøg** med anvendelse af 6 biler og 6 dage som blokke.

Fordele og ulemper ved planen
Fordelen ved en sådan plan er, at støj der kan skyldes eventuelle forskelle mellem biler og mellem dage elimineres. Ulempe er, at man let kommer op på mange delforsøg (her 36), og en ret kompliceret forsøgsplan. Det er således vigtigt at man foretager en randomisering, og det kan også være ret besværligt.

Eksempel 14.8 (romersk kvadratforsøg)
Der ønskes undersøgt hvorledes udbyttet af et biprodukt ved en fabrikationsproces afhænger af fabrikationsmetode og katalysatortype. Der skal vælges mellem 2 forskellige fabrikationsmetoder, og 3 katalysatorerer. Det kan ikke udelukkes, at de to faktorer vekselvirker, så man vælger som forsøgsplan en fuldstændig faktorstruktur.
Lad A_i betegne fabrikationsmetode i, og lad B_j betegne katalysator j. Der er følgelig 6 behandlinger $A_1 B_1$, $A_1 B_2$,$A_1 B_3$, $A_2 B_1$,$A_2 B_2$, $A_2 B_3$.
Hvert delforsøg varer en uge, og for at forsøgene ikke skal tage for lang tid, er man nødt til at anvende flere apparater samtidig. Da det kan tænkes, at såvel fabrikationstidspunkt som apparatvalg kan påvirke udbyttet, og dermed give en for stor spredning, vælges som forsøgsplan et romersk kvadratforsøg. Da der er 6 behandlinger bliver det et 6×6 kvadrat.
Forsøgsplan og forsøgsresultater var følgende:

		Apparater											
		1		2		3		4		5		6	
Uger	1	$A_2 B_1$	41.3	$A_2 B_2$	52.6	$A_1 B_2$	39.7	$A_2 B_3$	40.8	$A_1 B_3$	27.6	$A_1 B_1$	29.5
	2	$A_1 B_1$	9.3	$A_1 B_3$	18.7	$A_2 B_3$	19.5	$A_1 B_2$	21.1	$A_2 B_2$	31.9	$A_2 B_1$	33.2
	3	$A_2 B_3$	30.5	$A_2 B_1$	41.2	$A_1 B_3$	28.7	$A_2 B_2$	32.4	$A_1 B_1$	31.9	$A_1 B_2$	32.2
	4	$A_1 B_3$	50.7	$A_1 B_2$	48.3	$A_2 B_1$	51.6	$A_1 B_1$	40.9	$A_2 B_3$	60.4	$A_2 B_2$	49.8
	5	$A_2 B_2$	40.8	$A_2 B_3$	39.6	$A_1 B_1$	30.7	$A_2 B_1$	31.8	$A_1 B_2$	29.6	$A_1 B_3$	39.8
	6	$A_1 B_2$	42.7	$A_1 B_1$	52.4	$A_2 B_2$	62.1	$A_1 B_3$	53.6	$A_2 B_1$	63.6	$A_2 B_3$	58.7

Foretag en statistisk analyse af forsøget, idet man ønsker så lavt et udbytte som muligt.
Løsning:
Analyseres svarende til en firesidet variansanalyse uden vekselvirkning mellem blokke og faktorer.
Data indtastes

	metode	katalysator	uge	apparat	udbytte
1	a2	b1	u1	a1	41,3
2	a2	b2	u1	a2	52,6
3	a1	b2	u1	a3	39,7
4	a2	b3	u1	a4	40,8
5	a1	b3	u1	a5	27,6
6	a1	b1	u1	a6	29,5

Vælg Analyze ▶ Fit Model ▶ Indsæt for Y: udbytte ,Indsæt for ADD: metode, katalysator,apparat, uge ▶ Indsæt for CROSS: metode, katalysator (marker begge) ▶ Emphasis's rullemenu: vælg Minimal Report ▶ Run

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
metode	1	1	662,2044	21,9115	0,0001*
katalysator	2	2	27,8956	0,4615	0,6369
uge	5	5	4385,5389	29,0224	<,0001*
apparat	5	5	181,1189	1,1986	0,3454
metode*katalysator	2	2	61,4756	1,0171	0,3796

Da P-værdi for metode*katalysator =0.3796 > 0.05 ses heraf, at der ikke er vekselvirkning mellem metode*katalysator

Vekselsvirkningen "pooles" ned i "error"".

Marker under Source: metode*katalysator ▶ Remove

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
metode	1	1	662,2044	21,8775	0,0001*
katalysator	2	2	27,8956	0,4608	0,6367
uge	5	5	4385,5389	28,9774	<,0001*
apparat	5	5	181,1189	1,1967	0,3431

Det ses, at faktor katalysator ingen rolle spiller, mens metoden har stor betydning

Det ses endvidere, at det var fornuftigt at foretage en blokopdeling efter uge, mens det var unødigt at dele op efter apparater.

For at finde hvilken metode der giver det mindste udbytte

"Effect Details" ▶ rød pil ved metode ▶ Vælg "LS means student,s -t"

Level		Least Sq Mean
a2	A	43,433333
a1	B	34,855556

Levels not connected by same letter are significantly different.

Det ses, at metode 1 giver det laveste udbytte

Da størrelsen af udbyttet afhænger af blokken har det ingen mening at sige, at udbyttet er 34.85

Man kan naturligvis godt undersøge grafisk om data er normalfordelt, men da der ikke er gentagelser kan forudsætningen om varianshomogenitet ikke undersøges.

14.8. OVERSIGT OVER FREMGANGSMÅDE VED TOSIDET VARIANSANALYSE

Givet følgende skema

		Søjlefaktor Q			
		Q_1	Q_2	...	Q_q
Rækkefaktor R	R_1	x_{111} x_{112} .... x_{11n}	x_{121} x_{122} .... x_{12n}		x_{1q1} x_{1q2} .... x_{1qn}
	R_2	x_{211} x_{212} .... x_{21n}	x_{221} x_{222} .... x_{22n}		x_{2q1} x_{2q2} .... x_{2qn}
	...				
	R_r	x_{r11} x_{r12} .... x_{r1n}	x_{r21} x_{r22} .... x_{r2n}		x_{rq1} x_{rq2} .... x_{rqn}

$H_0: R*C = 0$ (Ingen signifikant vekselvirkning) H: $R*C \neq 0$

Data indtastes i 3 søjler R, Q og data.

Vælg Analyze ▶ Fit Model ▶ Indsæt for Y: data ,Indsæt for ADD: R og Q ▶ Indsæt for CROSS: R og Q (marker begge) ▶ Emphasis's rullemenu: vælg Minimal Report ▶ Run Model

(Minimal report er kun valgt for at undgå nogle i denne forbindelse overflødige figurer)

Der fremkommer en lang række tal.

Under "Effects Tests" findes P-værdi for vekselvirkning R*Q.

1) P - værdi < α : H_0 forkastes, dvs. Faktorerne R og C vekselvirker

Konfidens og d intervaller opstilles til afgørelse af hvilken kombination, idet man af faktorer, der eksempelvis giver mindst resultat.

"Effect Details" ▶ rød pil ved "R x Q ▶ Vælg "LS means Plot"

Der fremkommer en udskrift og en tegning.

Muligvis kan man ud fra tegning se hvilken kombination der eksempelvis giver lavest værdi.

Ellers kan det afklares ved på ovennævnte rullemenu at vælge "LSMeans students t"

Det giver en stor tabel (som kan fjernes ved med cursor på overskrift, højre musetast at fjerne markeringen ved "Crostab Report).

Under den findes en lille tabel hvor man kan se resultatet.

Ønskes fundet 95% konfidensintervaller

rød pil ved "karburator x olieblanding"▶ "LS Means Table" ▶Højremustast i tabel: Vælg "Columns" ▶vælg "lower" og derefter "upper".

2) P - værdi > α : H_0 accepteres, dvs, der kan ikke påvises en signifikant vekselvirkning

Vekselsvirkningen "pooles" ned i "error"".

Blandt mange tabeller findes under "Effect Tests" P-værdier for faktorerne R og C

a) Er de begge < α har begge en virkning

Under faktoren R kan man som i forrige tilfælde vælge "LSMeans students t" og finde hvilket niveau der giver det bedste resultat.

Samme gøres under den anden faktor

Et estimat for bedste værdi fås ved

Rød pil ved Response Profiler ▶ Factor profiling ▶ Profiler

Ved de fremkomne figurer flyttes linier, så man får bedst resultat.

Rød pil ved Predicter Profiler ▶ Confidence Intervals

Man kan nu beregne konfidensintervallerne

b) Kun en P-værdi < α .

Beregningerne ganske analoge med forrige tilfælde.

OPGAVER

Opgave 14.1

I et forsøg undersøgtes, om det kemiske udbytte af en proces afhænger af hvilken af 2 katalysatorer, der anvendes. Endvidere kan man benytte 3 forskellige apparater, og de kunne også tænkes at have indflydelse på resultatet.
Der fandtes følgende udbytter:

	Katalysator K_1	Katalysator K_2
Apparat A_1	69 72	65 66
Apparat A_2	72 71	70 69
Apparat A_3	70 71	72 73

1) Idet det antages at forudsætningerne for at udføre en variansanalyse er tilstede, skal der udføres en test til vurdering af, om middeludbyttets (eventuelle) afhængighed af de benyttede katalysatorer og apparater kan beskrives ved en additiv model.

2a) Hvis man af økonomiske grund vælger apparat A_1 hvilken katalysator skal man så vælge?

2b) Hvis man af økonomiske grunde vælger katalysator 1 hvilket apparat skal man så vælge?

2c) Hvilken (hvilke) kombinationer af apparat og katalysator giver det største udbytte.

Opgave 14.2

Man ønsker at undersøge den virkning som 2 faktorer (typen af glas og fosfor) har på skarpheden af billedet på en TV-skærm. Responsvariablen er den strøm (i microampere) som er nødvendig for at opnå et specifik skarpheds niveau.
Data er vist i nedenstående tabel:

		Fosfortype		
		1	2	3
Glastype	1	280 290 285	300 310 295	290 285 290
	2	230 235 240	255 240 235	220 225 230

Idet de sædvanlige variansanalyseforudsætninger antages opfyldt, ønskes følgende spørgsmål belyst:

1) Har glastype og fosfortype indflydelse på skarpheden?

2) Ud fra svaret i spørgsmål 2 skal angives, hvilken glastype og fosfortype der giver den største skarphed (giver den mindste respons)

Opgave 14.3

På en fabrik for glasvarer ønsker man at undersøge hvilken blandt 3 typer lim, der er bedst ved sammenlimning af 3 forskellige glastyper. Forsøget foregik ved, at man limede to glasplader sammen, og efter en passende tid undersøgte, hvor stor en kraft der skulle til for at trække pladerne fra hinanden. Man valgte at lave et fuldstændigt faktorforsøg med 5 gentagelser af hver behandling. Resultatet af forsøget var:

	Glastype A	Glastype B	Glastype C
LIM I	20 18 23 22 23	21 27 24 20 18	17 23 18 21 25
LIM II	30 25 28 27 28	28 24 16 25 21	28 25 29 27 28
LIM III	31 32 18 30 21	18 30 18 32 31	23 24 19 22 24

1) Kontroller om betingelserne om normalitet og varianshomogenitet er rimelig opfyldt.

Idet de sædvanlige variansanalyseforudsætninger antages opfyldt, ønskes følgende spørgsmål belyst:

2) Angiv hvilke faktorer der har en virkning.
3) Angiv den eller de kombinationer af type lim og type glas, der har den største sammenhængskraft. Angiv et 95% konfidensinterval for de pågældende kombinationer.

Opgave 14.4

Fabrikationen af et kemikalium baseres på en bestemt kemisk proces, som forudsætter tilsætning af katalysator og en PH - værdi på ca. 5.

Som led i en laboratoriemæssig undersøgelse af mulighederne for at forøge procesudbyttet foretoges bl.a. et forsøg, hvor man dels sammenlignede virkningen af tilsætning af 3 forskellige katalysatorer, dels undersøgte, om udbyttet afhang af, om den nødvendige PH - værdi opnåedes ved tilsætning af HCl i stedet for som hidtil H_2S0_4.

Forsøgsresultaterne var (udbytteprocenter):

		Tilsat syre			
		HCl		H_2S0_4	
Katalysatorer	1	27.0	27.7	30.1	29.1
	2	25.5	27.0	30.6	28.9
	3	26.5	25.0	25.8	28.0

1) Foretag en statistisk analyse af forsøgsresultaterne og drag konklusioner.
2) Estimer under hensyn til resultatet af den under punkt 1) foretagne analyse procesudbyttet ved benyttelse af katalysator 2 under tilsætning af HCl og opstil et 95% - konfidensinterval for dette udbytte.

Opgave 14.5

I et fremstillings proces, er en plastikstang fremstillet ved opvarmning af granuleret plastik til en smeltet tilstand og derefter samlet under tryk (pound per square inches) gennem en dyse. For at undersøge hvorledes disse 2 faktorer påvirker den samlingshastighed (målt i inches per. second) hvormed plastikstangen når tilstrækkelig styrke, laves et fuldstændigt faktorforsøg.

Tre slags granuleret plastik anvendes, og man vælger at benytte dem som en blokke i forsøget

Resultaterne blev

		blok					
		1		2		3	
		tryk		tryk		tryk	
		40	60	40	60	40	60
temperatur	200^0	1.35	1.74	1.31	1.67	1.40	1.86
	300^0	2.48	3.63	2.29	3.30	2.14	3.27

Foretag en statistisk analyse af forsøgsresultaterne og drag konklusioner.

Opgave 14.6

Hver af tre laboranter har bestemt hydroquinons smeltepunkt (^{0} Celcius) med (de samme) 4 termometre. Resultaterne var:

		Termometre			
		1	2	3	4
Laboranter	1	174.0	173.0	171.5	173.5
	2	173.0	172.0	171.0	171.0
	3	173.5	173.0	173.0	172.5

Det antages, at de nødvendige variansanalyseforudsætninger er opfyldt, og at termometre og laboranter ikke vekselvirker.

Følgende 2 spørgsmål ønskes belyst:

1. Aflæser laboranterne termometrene på samme måde?
2. Viser termometrene ens.

Opgave 14.7.

Ved en undersøgelse af, hvorledes virkningen af forskellige giftstoffer kunne bekæmpes, foretoges et fuldstændigt randomiseret forsøg, hvorved 2 giftstoffer og 4 vitaminbehandlinger inddroges i undersøgelsen, og overlevelsestiden (timer) af de benyttede forsøgsdyr måltes.

Nedenfor er anført en skematisk oversigt over forsøgsresultaterne:

		Vitaminbehandlinger			
		1	2	3	4
Giftstoffer	1	3.1 4.5	8.2 11.0	4.3 4.5	4.5 7.1
	2	2.2 2.1	3.0 3.7	2.3 2.5	3.0 3.6

Teoretiske overvejelser i forbindelse med tidligere lignende forsøg har vist, at variabeltransformationen $Y = \frac{1}{X}$ sikrer den for analysen nødvendige varianshomogenitet.

1) Analyser forsøgsresultaterne og drag konklusioner med hensyn til faktorernes virkemåde.
2) Opstil et 95%-konfidensinterval for den gennemsnitlige middeloverlevelsestid for hver enkelt vitaminbehandling og bestem den vitaminbehandling. som må antages at have bedst virkning.

Opgave 14.8.

Ved en enzymproduktion havde man valget mellem anvendelse af 4 forskellige råmaterialer R_1 R_2, R_3 og R_4. Man ønskede ved et forsøg at foretage en sammenligning mellem disse med hensyn til det opnåede procesudbytte.
Hvert delforsøg krævede en laborants fuldtidsmedvirken og varede 5 dage. Af ressourcernæssige grunde måtte forsøgsarbejdet fordeles på flere uger, idet man kunne få 4 laboranter stillet til disposition. Man kunne på forhånd ikke udelukke, at forsøgsresultaterne kunne afhænge af forsøgstidspunktet og af, hvilken laborant der medvirkede.
Forsøgsplan og forsøgsresultater var (kodede tal):

		Laboranter							
		1		2		3		4	
Uger	1	R_1	58	R_4	108	R_2	75	R_3	76
	2	R_4	60	R_3	26	R_1	-15	R_2	20
	3	R_3	-16	R_2	23	R_4	24	R_1	-24
	4	R_2	17	R_1	31	R_3	17	R_4	41

1) Foretag en statistisk analyse til undersøgelse af, om de 4 råmaterialer i middel giver samme udbytte
2) Find i benægtende fald den type råmateriale der giver størst udbytte.
3) Undersøg tillige, om procesudbyttet i middel må antages at være påvirket af forsøgstidspunkt og/eller laborant.

Opgave 14.9.

Ved en tekstilfabrikation måltes for to forskellige vævemetoder og 5 forskellige materialetyper antallet af garnbrud pr. 1000 m^2 klæde. Resultaterne var følgende:

		Materialetyper				
		1	2	3	4	5
Vævemetoder	1	4	12	23	6	9
	2	4	3	7	2	3

Foretag efter en passende variabeltransformation en analyse af, om og i bekræftende fald hvorledes middelantallet af garnbrud afhænger af vævemetoder og/eller materialetyper.

15. ENKELT REGRESSIONSANALYSE

15.1 INDLEDNING

I dette kapitel betragtes forsøg, hvor man har målt sammenhørende værdier af to **kvantitative** variable. Det følgende eksempel demonstrerer et sådant tilfælde.

Eksempel 15.1.Kvantitative variable

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

Mængden af kunstfibre x er blevet bestemt på forhånd (har fået ganske bestemte værdier), så den er ikke en statistisk variabel. Trækstyrken Y synes derimod udover mængden af kunstfibre også at være påvirket af andre ukendte og ukontrollable "støjfaktorer". Y må derfor opfattes som en statistisk variabel.
I andre situationer er både X og Y statistiske variable. Dette gælder eksempelvis, hvis man ønsker at undersøge om der er en sammenhæng mellem personers højde Y og vægt X, og derfor for en række personer måler sammenhørende værdier af højde og vægt.

Målet med en regressionsanalyse er at finde en funktionssammenhæng mellem den uafhængige variabel y og de afhængige variable.

I eksempel 15.1 ville man umiddelbart sige, at da man har 15 punktpar, så vil et polynomium af fjortende grad $y = a_{14}x^{14} + a_{13}x^{13} + \ldots + a_1x + a_0$ gå igennem alle punkter, og det må derfor være en god model. Dette er imidlertid ikke tilfældet, da y - værdierne jo er resultater af forsøg der er påvirket af ukontrollable støjkilder. Polynomiets koefficienter vil derfor afspejle disse tilfældige udsving, og det giver derfor en ganske meningsløs model. Endvidere er modellen alt for matematisk kompliceret til at kunne bruges i praksis. Vi søger derfor i regressionsanalysen en enklere model, som tager rimeligt hensyn til støjen ved målingerne.

Er den ene variabel som i eksempel 15.1 en (kontrolleret) ikke statistisk variabel, så har man mulighed for hver x- værdi, at foretage gentagne målinger af den statistiske variabel Y (randomiseret). Dette giver mulighed for at beregne et estimat for den spredning der skyldes støjen, hvilket (som det vises i afsnit 15.2.3) kan udnyttes ved testning af den foreslåede model.

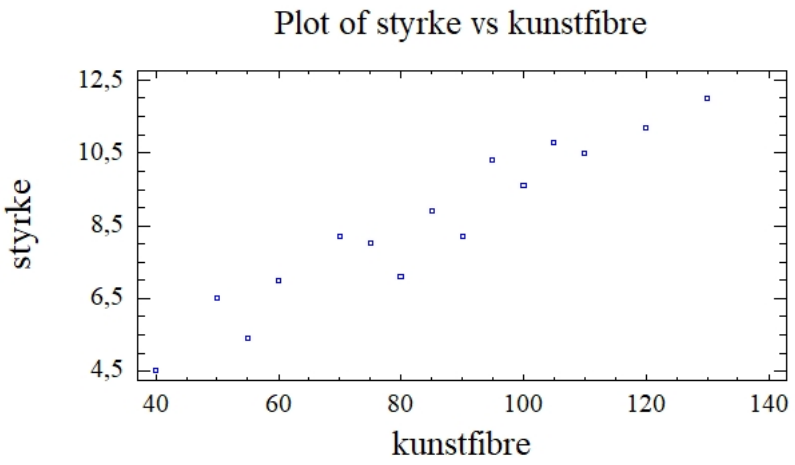
Lineær model.
Ved en lineær model forstås her en model, der er lineær med hensyn til parametrene.
Eksempelvis er såvel funktionen $y = f(x) = a + bx + cx^2$ som $y = g(P,T) = a + b \cdot P + c \cdot T$ lineære i de 3 parametre a, b og c . Som et eksempel på en model der ikke er lineær i parametrene kan nævnes $y = a + bx^c$.

Vi vil i dette kapitel betragte det ved anvendelserne meget ofte forekomne tilfælde, som kaldes **enkelt regressionsanalyse**, og hvor modellen er lineær i 2 parametre, eller kan gøres lineær ved en passende transformation..
Som eksempler herpå kan nævnes $y = a + bx$ og $\ln y = a + b \cdot \ln x$.

15.2. BESTEMMELSE AF REGRESSIONSLIGNING

Vi betragter igen eksempel 15.1

Regressionslinie og regressionskoefficienter.
Afsættes 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).
Derfor er det nærliggende at antage, at **middelværdien** af den statistiske variable *Y* er en lineær funktion af *x* af formen $E(Y|x) = \beta_0 + \beta_1 x$. (1)

$E(Y|x)$ skal læses middelværdien af *Y* for fastholdt *x*.

Vi vil i det følgende ofte i ligningen (1) kort skrive *Y* eller μ fremfor $E(Y|x)$.

Ligning (1) kaldes **regressionsligningen** (eller den teoretiske regressionsligning), grafen kaldes for **regressionslinien** (eller den teoretiske regressionslinie) , og konstantledet β_0 og hældnings-koefficienten β_1 kaldes **regressionskoefficienterne**.

Mens middelværdien af Y ligger på regressionslinien, kan den aktuelle observerede værdi af Y ikke forventes at ligge på den.
For et punktpar (x_1, y_i) gælder derfor, at $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, hvor ε_i kaldes den i'te residual.

Bestemmelse af regressionslinien ved mindste kvadraters metode.

På basis af en række sammenhørende værdier af x og y bestemmes estimater $\hat{\beta}_0$ og $\hat{\beta}_1$ for regressionskoefficienterne β_0 og β_1 ved " mindste kvadraters metode". Værdierne $\hat{\beta}_0$ og $\hat{\beta}_1$ kaldes de empiriske regressionskoefficienter.
Kan det ikke misforstås, så kort blot regressionskoefficienterne.

Det følgende eksempel viser metoden anvendt på et lille taleksempel. Man benytter hertil et matrixprogram. De angivne metoder kan umiddelbart generaliseres til mere komplicerede eksempler.

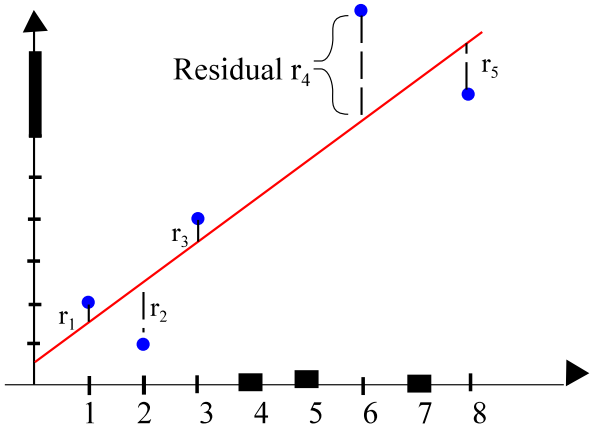
Eksempel 15.2. Bestemmelse af regressionskoefficienter ved mindste kvadraters metode.
I et medicinsk forsøg måles på en forsøgsperson sammenhørende værdier af en bestemt medicin i blodet (i %) og reaktionstiden. Resultaterne var:

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

Bestem ved mindste kvadraters metode et estimat for regressionslinien. ◆

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

På figur 15.1 er afsat de 5 punkter, og indtegnet en ret linie.



Figur 15.1 *Residualer*

Mindste Kvadraters metode. Regressionslinien $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ bestemmes som den af alle mulige rette linier, for hvilket summen af kvadratet af residualerne til linien er mindst.
I eksempel 15.2 er kvadratsummen $r_1^2 + r_2^2 + r_3^2 + r_4^2 + r_5^2$.
Løsningen af dette optimeringsproblem er angivet nedenfor.

Bestemmelse af regressionsligningen ved mindste kvadraters metode
LØSNING:

I vort tilfælde hvor vi har 5 punkter, indsættes vi disse i ligningen $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Dette giver: $2 = \hat{\beta}_0 + \hat{\beta}_1 \cdot 1,\quad 1 = \hat{\beta}_0 + \hat{\beta}_1 \cdot 2,\quad 4 = \hat{\beta}_0 + \hat{\beta}_1 \cdot 3,\ 9 = \hat{\beta}_0 + \hat{\beta}_1 \cdot 6,\ 7 = \hat{\beta}_0 + \hat{\beta}_1 \cdot 8$.

De 5 ligninger med 2 ubekendte $\hat{\beta}_0$ og $\hat{\beta}_1$ kan i matrixnotation skrives:

$$Y = X \cdot B \text{ hvor } Y = \begin{pmatrix} 2 \\ 1 \\ 4 \\ 9 \\ 7 \end{pmatrix},\quad X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 6 \\ 1 & 8 \end{pmatrix} \text{ og } B = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$$

De søgte værdier af $\hat{\beta}_0$ og $\hat{\beta}_1$ findes som den løsning til dette overbestemte ligningssystem som giver den mindste RMS - fejl.
Løsningen er (se evt. "Matematik for ingeniører" bind 3 eller Matricer og lineære ligninger afsnit 10) bestemt ved løsningen $B = (X^T \cdot X)^{-1} \cdot X^T \cdot Y$.
I vort taleksempel er

$$B = \left(\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 6 & 8 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 6 \\ 1 & 8 \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 6 & 8 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \\ 4 \\ 9 \\ 7 \end{pmatrix} = \begin{pmatrix} 0.6 \\ 1.0 \end{pmatrix}$$

Regressionsligningen bliver følgelig $\underline{\underline{\hat{y} = 0.6 + 1.0x}}$

I praksis vil man benytte et færdigt program til bestemmelse af regressionskoefficienterne. Dataene indtastes enten i et statistikprogram som SAS.JMP, eller i en lommeregner med regressionsprogram som TI-89 eller TI-Nspire.

15.3 VURDERING AF OM REGRESSIONSLIGNING BESKRIVER DATA GODT.

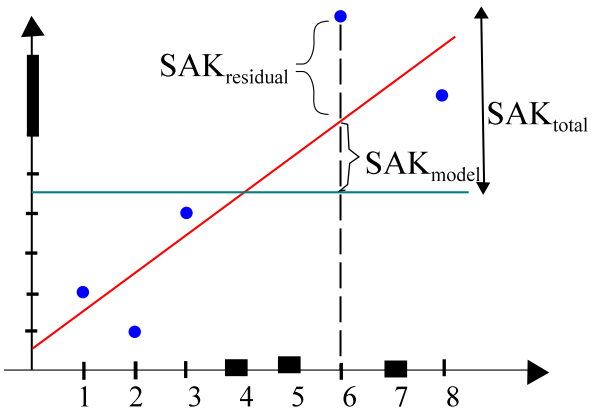
Vi vil i dette afsnit se på det tilfælde, hvor der for hver x - værdi kun er målt én y - værdi.

Det er altid muligt ved mindste kvadraters metode at finde en sådan "mindste kvadraters linie". Det betyder ikke nødvendigvis, at linien så også er en rimelig model, som kan anvendes til at beskrive sammenhængen.

Til vurdering heraf vil man
1) se på en **tegning**.
"Mindste kvadraters linie" tegnes i et koordinatsystem sammen med punkterne. Hvis den lineære model beskriver dataene godt, skal punkterne fordeler sig tilfældigt omkring linien
I mere komplicerede tilfælde, er det nødvendigt i stedet at afsætte residualerne (i et såkaldt residualplot). Residualerne bør så fordele sig tilfældigt omkring den vandrette 0 - linie

2) Se på **størrelsen af forklaringsgraden** r^2 (også kaldet determinationskoefficient).
   Den angiver et talmæssigt mål for hvor tæt punkterne ligger på linien.
   Sædvanligvis finder man, at den fundne model på tilfredsstillende måde beskriver data, hvis forklaringsgraden er på over 70% samtidig med, at tegningen viser, at punkterne fordeler sig tilfældigt omkring den fundne regressionskurve.
   Selv om forklaringsgraden er lav, kan der dog godt være en vis afhængighed mellem de to variable.
   Eksempelvis, hvis man måler sammenhørende værdier af højde og vægt for alle værnepligtige mænd, vil forklaringsgraden være lav (under 50%). Ser man på den såkaldte korrelationskoefficient ρ vil den dog være positiv.
   Hvis det ved en test viser sig at være signifikant større end 0 (svarer til at man tester om liniens hældningskoefficient er større end 0) , har vi vist, at i middel vil der gælde, at jo højere man er jo mere vil man i middel veje.
   Målepunkterne ligger imidlertid så spredt, at dette blot afspejler en tendens. Man kan ikke benytte regressionslinien til at forudsige noget med blot en rimelig sikkerhed.
3) Undersøge om der er **outliers**.
   Outliers er en betegnelse for enkelte målinger som afviger så kraftigt fra den almindelige tendens, at det kan skyldes fejlmålinger. Sådanne punkter kan i uheldige tilfælde på grund af et stort bidrag til residualsummen få regressionslinien til at dreje. Det er dog klart, at man ikke blot kan stryge sådanne "ubehagelige" punkter. Det må kun ske, hvis man er sikker på, at punktet skyldes en fejl af en eller anden art ved målingen.

Definition og beregning af forklaringsgrad.
I praksis vil lade en lommeregner eller en PC med et statistikprogram beregne de enkelte statistiske størrelser.
Ved tolkningen af de fremkomne størrelser vil en anskuelig forståelse af størrelserne dog være nyttig.
I det følgende vil vi derfor definere nogle fundamentale definitioner, og søge at anskueliggøre dem dels på figur 15.3 dels ved at foretage beregningeren på tallene i eksempel 15.2.



Figur 15.3. SAK - størrelser

Definitioner:

SAK_{total} = sum af kvadrater af residualerne til den vandrette linie $y = \bar{y}$

Data i eksempel 15.2 giver: $\bar{y} = \frac{2+1+4+9+7}{5} = 4.6$

De 5 punkters "residualer" til den vandrette linie $y = \bar{y}$.

$r_1 = 2 - 4.6 = -2.6, r_2 = 1 - 4.6 = -3.6, r_3 = 4 - 4.6 = -0.6, r_4 = 9 - 4.6 = 4.4, r_5 = 7 - 4.6 = 2.4$.

Vi får $SAK_{total} = r_1^2 + r_2^2 + \ldots + r_5^2 = 2.6^2 + 3.6^2 + 0.6^2 + 4.4^2 + 2.4^2 = 45.2$.

$SAK_{residual}$ = sum af kvadrater på de enkelte punkters afstand fra den fundne regressionslinie .

Af eksempel 15.2 fås følgende residualer til den fundne regressionslinie $y = 0.6 + 1.0\,x$:

$r_1 = 2 - (0.6 + 1 \cdot 1) = 0.4, r_2 = 1 - (0.6 + 1 \cdot 2) = -1.6, r_3 = 4 - (0.6 + 1 \cdot 3) = 0.4,$

$r_4 = 9 - (0.6 + 1 \cdot 6) = 2.4, r_5 = 7 - (0.6 + 1 \cdot 8) = 0.4$.

$SAK_{residual} = r_1^2 + r_2^2 + \ldots + r_5^2 = 0.4^2 + (-1.6)^2 + 0.4^2 + 2.4^2 + .(-1.6)^2 = 11.2$

SAK_{model} = sum af kvadrater af "regressionsliniens afstand" fra det totale gennemsnit $\bar{y}$.

Af eksempel 15.2 fås residualerne for regressionslinien $y = 0.6 + 1.0\,x$'s "afstand" fra det totale gennemsnit $\bar{y}$=4.6.

$r_1 = 0.6 + 1 \cdot 1 - 4.6 = -3.0, r_2 = 0.6 + 1 \cdot 2 - 4.6 = -2.0, r_3 = 0.6 + 1 \cdot 3 - 4.6 = -1.0,$

$r_4 = 0.6 + 1 \cdot 6 - 4.6 = 2.0, r_5 = 0.6 + 1 \cdot 8 - 4.6 = 4.0$.

$SAK_{model} = r_1^2 + r_2^2 + \ldots + r_5^2 = (-3.0)^2 + (-2.0)^2 + (-1.0)^2 + 2.0^2 + .4.0^2 = 34$

Der gælder generelt, at **$SAK_{total} = SAK_{residual} + SAK_{model}$** (jævnfør, at 45.2 = 11.2 + 34)

Forklaringsgraden r^2 er bestemt ved

$$r^2 = \frac{SAK_{model}}{SAK_{total}} = \frac{SAK_{total} - SAK_{residual}}{SAK_{total}} = 1 - \frac{SAK_{residual}}{SAK_{total}} \quad (r^2 = \frac{34}{45.2} = 0.752)$$

<u>Anskuelig forklaring:</u>

Ligger punkterne tæt ved linien, er residualerne små, og $SAK_{residual}$ lille.

Nu er lille jo et relativt begreb, så man sammenligner med SAK_{total}, dvs. med residualerne til den vandrette linie $y = \bar{y}$. Hvis der ingen sammenhæng er mellem y- og x- værdierne (*Y* er uafhængig af *x*) vil regressionslinien stille sig næsten vandret, dvs. $y \approx \bar{y}$.

Måler man eksempelvis sammenhørende værdier af personers højde, og deres IQ (intelligenskvotient) vil man uden tvivl få en ret linie der er praktisk taget vandret, og en forklaringsgrad der ligger tæt på 0.

Det betyder igen at $SAK_{residual} \approx SAK_{total}$ og dermed at $r^2 = 1 - \frac{SAK_{residual}}{SAK_{total}} \approx 0$.

Hvis derimod der er en sammenhæng *(Y* er lineært afhængig af *x*) vil regressionslinien have en hældning forskellig fra nul. Det betyder igen at $SAK_{residual} << SAK_{total}$.

Heraf følger, at $r^2 \approx 1$.

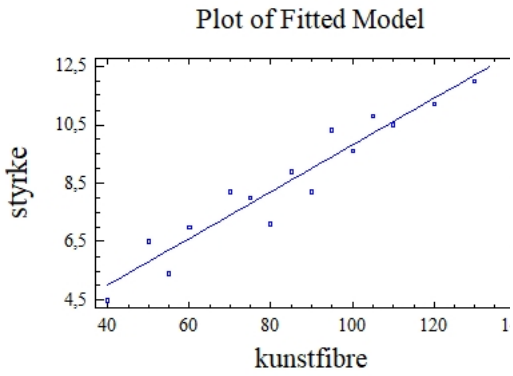
Man siger også, at den fundne model "forklarer " $r^2 \cdot 100\%$ af den "totale variation"

I eksempel 15.2 forklarer den fundne model således " 75.2% af den totale variation.

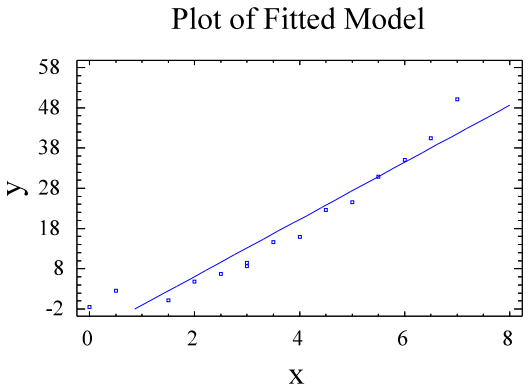
Transformation. Hvis man ikke finder, at en ret linie beskriver data godt nok, så er det jo muligt, at en anden kurve bedre beskriver sammenhængen. Eksempelvis er det jo velkendt fra matematikken, at graferne for eksponentialfunktioner og potensfunktioner ved en passende logaritmisk transformation kan blive til rette linier. Det giver naturligvis lidt mere komplicerede regninger, men statistikprogrammer og også en del lommeregnere kan dog let foretage en regressionsanalyse også i sådanne tilfælde. I eksempel 15.5 er et sådant eksempel gennemgået.

At man ikke alene kan stole på forklaringsgraden illustreres ved følgende eksempel.

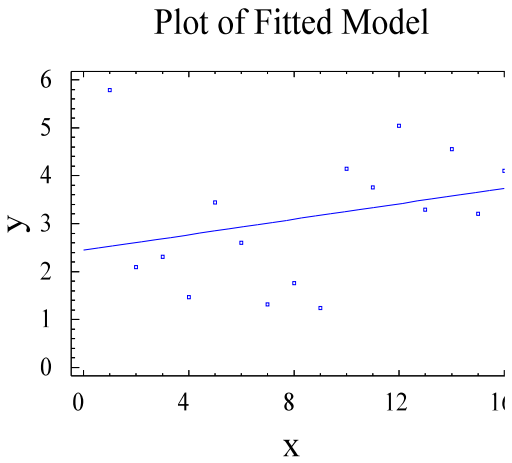
Eksempel 15.3 .Grafisk vurdering af model.
De følgende 4 figurer afspejler forskellige muligheder.



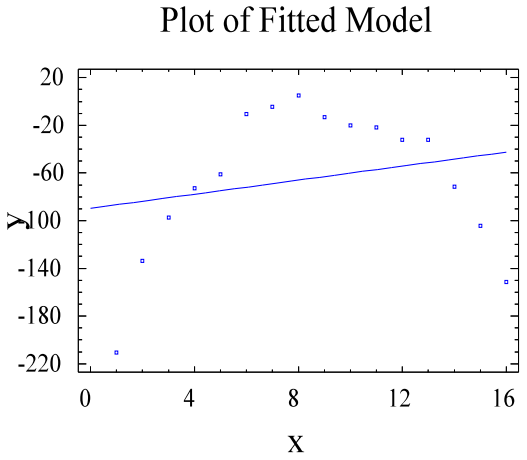
Figur 15.2a: $r = 0.959$ $r^2 = 91.9\%$



Figur 15.2b: $r = 0.962$ $r^2 = 92.6\%$



Figur 15.2c: $r = 0.278$ $r^2 = 7.73\%$



Figur 15.2d: $r = 0.229$ $r^2 = 5.24\%$

I figur 15.2a synes den lineære model at kunne beskrive dataene godt, idet punkterne fordeler sig tilfældigt omkring linien, og forklaringsgraden $r^2 = 91.9\%$ er høj.

I figur 15.2b er forklaringsgraden også høj, og punkterne ligger da også tæt ved linien. Imidlertid ligger punkterne ikke tilfældigt omkring linien. Yderpunkterne ligger over og de midterste punkter under linien, så det er næppe rimeligt at anvende en ret linie som model. I stedet kunne man overveje en eksponentialfunktion eller et andengradspolynomium.

I figur 15.2c er regressionslinien er næsten vandret, og forklaringsgraden ringe.
En test må afgøre om hældningen er signifikant forskellig fra 0. Er dette ikke tilfældet, er der ingen relation mellem x og y (de er uafhængige)y.
Er x og y uafhængige vil punkterne fordele sig tilfældigt omkring gennemsnitslinien $y = \bar{y}$, og forklaringsgraden være 0.
I figur 15.2d er forklaringsgraden også lille, men alligevel må vi antage at der er en sammenhæng mellem x og y. Den er blot ikke lineær, men muligvis en parabel. ◆

Sammenhæng mellem korrelationskoefficient og forklaringsgrad.
Hvis både X og Y er normalfordelte statistiske variable (som eksempelvis når man aflæser sammenhørende værdier af højde og vægt for en række personer) angiver korrelationskoefficienten ρ en størrelse mellem -1 og 1 som kan anvendes til at angive, om der er en sammenhæng (korrelation) mellem X og Y. Er korrelationskoefficienten positiv har punkterne en voksende tendens, hvis den er negativ har de en aftagende tendens.
Et estimat for ρ er størrelsen r. Kvadreres den er r^2 den samme som forklaringsgraden.

Ekstrapolation. Selv om modellen synes på tilfredsstillende måde at beskrive data, så er det jo faktisk kun sikkert indenfor måleområdet. Man skal være yderst forsigtig med at ekstrapolere, dvs. på basis af modellen for x - værdier udenfor måleområdet beregne hvad y er.

Årsagssammenhæng. Selv om man finder, at der er en sammenhæng mellem x og y, er det ikke sikkert, at der er en årsagssammenhæng.
Der findes en god korrelation mellem antallet af storke i Sønderjylland i 1930-erne og antallet af børnefødsler (de faldt begge i samme takt), men det ene er nok ikke årsagen til det andet.
Man kender det også fra sammenhængen mellem kræft og tobaksrygning, hvor der i mange år var en diskussion om der ene bevirkede det andet, eller om det var en hel tredie faktor, der fik antallet af lungekræft til at stige.

Eksempel 15.4 Vurdering af model

Tilsætning af en vis mængde kunstfibre forøger et garns trækstyrke. 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 p. kg uld	40	50	55	60	70	75	80	85	90	95	100	105	110	120	130
Trækstyrke : 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

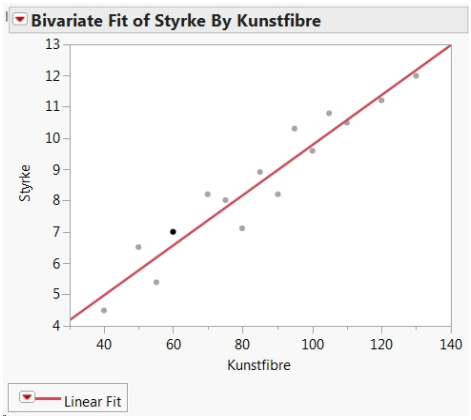
1) Find r^2 og anvend denne samt en tegning med indlagt ret linie (eller residualernes fortegn) til vurdering af modellen.
2) Opskriv regressionsligningen.

Løsning

1) Data indtastes

```
kunstfibre   styrke
40           4,5
50           6,5
osv.
```

Man kan ved analysen vælge 2 modeller, enten "fit Y by X" som giver en forholdsvis simpel og overskuelig analyse, eller "Fit model" som er nødvendig ved mere specielle analyser.
Vælg Analyze ▶ Fit Y by X ▶ markér "Styrke" og tryk på Y Response ▶ markér "kunstfibre" og tryk på X Factor ▶ OK
Der fremkommer et "scatterplot", hvor man i et koordinatsystem kan se punkterne afbildet.
Rød pil ▶ vælg fra rullemenuen "Fit Line".
Der fremkommer så følgende tegning og udskrift:
Af figuren ses, at punkterne fordeler sig tilfældigt omkring linien.



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

Linear Fit

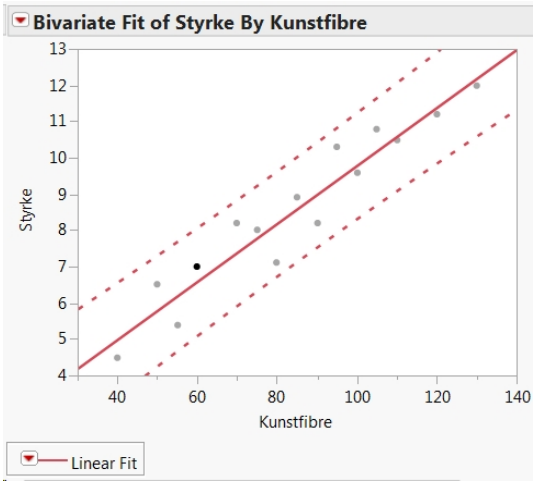
Styrke = 1,8086555 + 0,0798974*Kunstfibre

Summary of Fit

RSquare	0,919301
RSquare Adj	0,913093
Root Mean Square Error	0,648068
Mean of Response	8,546667
Observations (or Sum Wgts)	15

Outliers. Af ovenstående figur ses, at der næppe er nogen "outliers" (punkter der afviger så kraftigt fra det generelle billede, at man kunne frygte de var fejlmålinger).
En undersøgelse af om der er outliers er vigtigt.
En (lidt usikker) metode er, at få tegnet såkaldte 95% predikationskurver, og se om praktisk taget alle punkter ligger indenfor disse.

Rød pil under tegningen ved "linear fit" ▶ vælg"Confid. Curves indv"



Da alle punkter ligger indenfor grænserne, tyder det ikke på, at der er outliers.

En lidt sikrere metode er, at lade SAS-JMP beregne såkaldte "Studentized Residuals", som tager i betragtning, at spredningen er mindre ved "midtpunktet" end langt fra det.
Vælg Analyze▶Fit model▶markér "Styrke" og tryk på Y ▶markér "kunstfibre" og tryk ADD▶I "Emphasis" vælg "Minimal Report" (for at begrænse udskrifterne)▶Run

Vi får tegninger og udskrifter nogenlunde magen til før.
Rød pil▶vælg Save Columns▶Studentized residuals
Der bliver nu tilføjet en ekstra kolonne til data

Kunstfibre	Styrke	Studentized Resid Styrke
40	4,5	-0,910214041
50	6,5	1,1923782679
55	5,4	-1,347927857
60	7	0,6566305986
70	8,2	1,2900712404
75	8	0,3194449843
80	7,1	-1,759454913
85	8,9	0,479282925
90	8,2	-1,279096524
95	10,3	1,4483232937
100	9,6	-0,321237896
105	10,8	0,9851242793
110	10,5	-0,161471825
120	11,2	-0,33813566
130	12	-0,355350614

Heraf fremgår, at da ingen "Studentized Residuals", numerisk er større end 3 (det er tilladt, at nogle få er større end 2) er der ingen outliers.
Af fortegnene på Studentized Residuals kan man se at de har ret skiftende fortegn hvilket jo (måske lettere) viser at punkterne fordeler sig tilfældigt omkring linien.
Konklusion: Modellen synes tilstrækkelig godt at beskrive data indenfor måleområdet.

2) Under "Linear Fit" ses regressionsligningen
Styrke = 1.8087 + 0.0799*kunstfibre.

15.4 TEST OG KONFIDENSINTERVALLER

De foregående betragtninger kræver (bortset fra kontrol af Outliers) ingen statistiske forudsætninger, idet man jo altid ved mindste kvadraters metode kan beregne regressionskoefficienterne, beregne forklaringsgrad, tegne kurver og punkter ind i et koordinatsystem og så herfra vurdere om modellen er acceptabel.

Forudsætninger for regressionsanalyse.
Ønsker vi at foretage en nøjere statistisk analyse som eksempelvis at teste "om Y er uafhængig af x, dvs. af om $\beta_1 = 0$", eller opstille konfidensintervaller for β_1 må observationerne opfylde de samme krav som ved variansanalyser, dvs.

1) Uafhængige målinger:
   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.
   Kravet kan opfyldes ved en hensigtsmæssig forsøgsplan. I eksempel 15.2 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.
2) Residualerne skal være rimelig normalfordelt.
   Analysen er robust overfor afvigelser, men er man i tvivl kan man indtegne residualerne på et normalfordelingsplot.
3) Varianshomogenitet.
   Variansen af residualerne (eller af Y) Y skal være den samme uafhængig af x's værdi.
   Igen er analysen robust overfor afvigelser, så man vil sædvanligvis nøjes med at se på et residualplot, og deraf se, at residualerne fordeler sig jævnt uden store udsving.

Test af om *Y* er uafhængig af *x*
Lad os antage, at forudsætningerne er opfyldt, og at vi har fundet (ved at betragte tegning + forklaringsgrad og manglende outliers), at modellen $E(Y|x) = \beta_0 + \beta_1 x$ gælder.
Hvis Y er uafhængig af x betyder det, at regressionslinien er vandret, eller at hældningskoefficienten β_1 er 0.
Vi får altså: H_0: Y er uafhængig af $x \Leftrightarrow H_0$: Regressionslinien er vandret $\Leftrightarrow H_0: \beta_1 = 0$.

Metode 1: *F* - test.
Ifølge definitionen af SAK_{model} angiver denne kvadratsummen af punkter på regressionsliniens afstand til det totale gennemsnit $\bar{y}$ (se side 101)
Det er derfor klart, at hvis regressionslinien er næsten vandret, vil $SAK_{model} \approx SAK_{residual}$
Man kan vise, at $s^2_{model} = \frac{SAK_{model}}{1}$ har frihedsgraden 1 og $s^2_{residual} = \frac{SAK_{residual}}{N-1}$ har frihedsgraden N -1
Idet $F_{model} = \frac{s^2_{model}}{s^2_{residual}}$ kan vises, at en test for om linien er vandret er en sædvanlig F-test.
Metoden er forklaret mere udførligt i notatet "Supplement til Anvendt statistik" som findes på min hjemmeside.

H_0 forkastes, hvis P - værdi $= P(F > F_{model}) < \alpha$, hvor F_{model} er F- fordelt med en tællerfrihedsgrad på 1 og en nævnerfrihedsgrad på N - 2.

Metode 2. t - test.

Metoden er forklaret mere udførligt i notatet "Supplement til Anvendt statistik" som findes på min hjemmeside.

Lad $t = \dfrac{\hat{\beta}_1}{s_{\beta_1}}$, hvor $s_{\beta_1} = \dfrac{\left|\hat{\beta}_1\right| \cdot s_{\text{residual}}}{s_{\text{model}}}$ er et estimat for spredningen på β_1.

Det kan vises, at t et t - fordelt med N - 2 frihedsgrader.

H_0 forkastes, hvis P - værdi = $P(T > |t|) < \alpha$

På tilsvarende måde kan man teste $H_0: \beta_1 \geq 0$ og $H_0: \beta_1 \leq 0$ ved ensidede test.

Endvidere kan man analogt teste nulhypotesen H_0: $\beta = k$ hvor k er en konstant

Konfidensintervaller og prædistinationsintervaller.

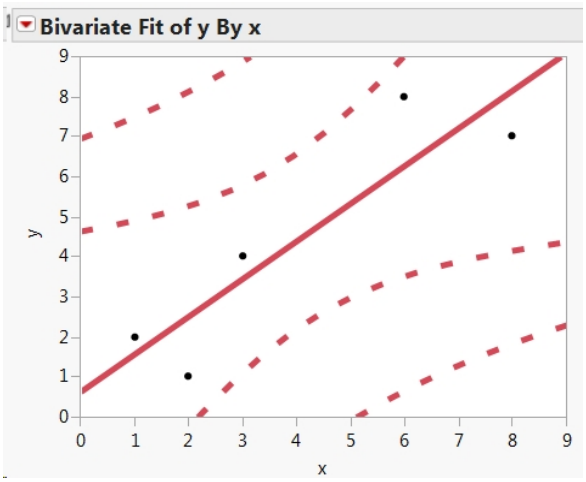
Et led i analysen kan være, at udregne et 95% konfidensinterval for β_1.

Endvidere vil man ofte være interesseret i en speciel værdi for x, for hvilken man ønsker beregnet såvel den tilsvarende "forventede" y - værdi (predicted value") som et 95% -konfidensinterval for middelværdien $\tilde{\mu}$ og et 95% prædistinationsinterval for én **ny** observation.

På figur 15.4 er her tegnet 95% konfidensintervaller for middelværdierne (de inderste buede kurver), og 95% prædistinationsintervaller (de yderste to kurver).

Man ser tydeligt, at konfidensintervallerne er smallest omkring "centrum" $(\bar{x}, \bar{y})$.

I notatet "Supplement til Anvendt statistik" som findes på min hjemmeside er angivet formler for bl.a. disse konfidensintervaller.



Figur 15.4 Konfidens- og prædistinations-intervaller

Eksempel 15.5 (fortsættelse af eksempel 15.4) Test

Tilsætning af en vis mængde kunstfibre forøger et garns trækstyrke. 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 p kg uld	40	50	55	60	70	75	80	85	90	95	100	105	110	120	130
Trækstyrke : 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

Det forudsættes, at forudsætningerne for regressionsanalyse er opfyldt.

I eksempel 15.4 fandt man at ligningen y = 1.8087+0.0799x var en god model for data.

1) Test om y er uafhængig af x
2) Find 95% konfidensinterval for hældningen β
3) Find den til x = 65 svarende værdi for y, samt et 95% konfidensinterval for y .
4) Find 95% prædistinationsinterval for 1 ny observation svarende til x - værdien 65.

Løsning:

Data er indtastet som i eksempel 15.4

1) $H_0: Y$ er uafhængig af $x \Leftrightarrow H_0:$ Regressionslinien er vandret $\Leftrightarrow H_0: \beta_1 = 0$.

Vælg Analyze ▶Fit model ▶markér "Styrke" og tryk på Y ▶markér "kunstfibre" og tryk på Add ▶Run

Der fremkommer så blandt andet denne udskrift:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	62,197436	62,1974	148,0919
Error	13	5,459897	0,4200	Prob > F
C. Total	14	67,657333		<,0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	1,8086555	0,578421	3,13	0,0080*
Kunstfibre	0,0798974	0,006565	12,17	<,0001*

Det ses, ud for "Model", at
F - Ratio = 148.09 og at P-value = 0.0001
(kan også ses ud for "kunstfibre"
Heraf fås, at
H_0 forkastes
Y er ikke uafhængig af x.

2) Konfidensinterval for hældningskoefficienten β_1:

Cursor placeres i tabel for "Parameter Estimates", højre musetast ▶Columns ▶"Lower 95%" ▶ "Upper 95%"

Man får bl.a. følgende tabel:

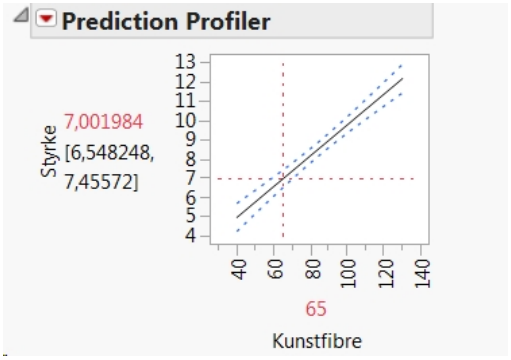
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	1,8086555	0,578421	3,13	0,0080*	0,5590522	3,0582587
Kunstfibre	0,0798974	0,006565	12,17	<,0001*	0,0657135	0,0940812

Heraf aflæses [0.0657 ; 0.0940]

3) 95% konfidensinterval for middeltrækstyrken svarende til x - værdien 65.
rød pil ved "Response, y ▶vælg "Factor Profiling"▶ Profiler ▶sæt cursor på tallet i bunden og skriv 65 og aflæs

x = 65 Y = 7.002
95% konfidensinterval =[6.55;7.46]



4) Indsæt 65 nederst i x-kolonne i tabel
rød pil ved response ▶save Columns ▶Individ Confidence Interval
I tabel fremkommer nu prædistinationsintervaller [5.53 ; 8.47]

15.5. TRANSFORMATION AF DATA.

Da man altid vil foretrække den simplest mulige model, er modellen $y = ax + b$ altid den, man starter med at anvende.
Hvis man ser, at punkterne ikke ligger tilfældigt omkring linien, men dog synes at følge en krum kurve, så må man anvende en anden model.

Blandt de oftest forekommende modeller er

1) Eksponentialmodellen $y = a \cdot b^x$
   Tages logaritmen på begge sider fås $\ln(y) = \ln(a) + \ln(b) \cdot x$
   som er en normal retlinet model, hvis man erstatter $\ln(y)$ med z
2) Potensmodellen $y = a \cdot x^b$
   Tages logaritmen på begge sider fås $\ln(y) = \ln(a) + b \cdot \ln(x)$
   som er en normal retlinet model, hvis man erstatter $\ln(y)$ med z og $\ln(x)$ med v.
3) Logaritmemodellen $y = a + b \cdot \ln(x)$
   som er en normal retlinet model, hvis man erstatter $\ln(x)$ med v.

Generelt gælder, at man så vidt mulig foretrækker modeller med kun 2 parametre a og b (som (1), (2) og (3)), fremfor modeller som eksempelvis andengradsmodellen $y = ax^2 + bx + c$ der indeholder 3 parametre, da de mest stabile (set fra et statistisk synspunkt).

I den følgende oversigt er angivet en liste med kommentarer over de mest almindelige transformationer.

Transformation til lineær model .

Nr	Model	Kommentar
1	Linear model: Y = a + b*X	
2	Exponential model: Y = exp(a + b*X)	$Y = e^{a+bX} = e^{a} \cdot e^{bX} \Leftrightarrow \ln(Y) = a + bX$ Sættes Z =ln(Y) fås Z= a+b · X
3	Reciprocal-Y model: Y = l/(a + b*X)	$Y = \frac{1}{a + b \cdot X} \Leftrightarrow \frac{1}{Y} = a + b \cdot X$ Sættes $Z = \frac{1}{Y}$ fås Z= a+b · X
4	Reciprocal-X model: Y = a + b/X	$Y = a + \frac{b}{X}$. Sættes $W = \frac{1}{X}$ fås Y= a+b · W
5	Double reciprocal model: Y = l/(a + b/X)	$Y = \frac{1}{a + \frac{b}{X}} \Leftrightarrow \frac{1}{Y} = a + b\frac{1}{X}$ Sættes $Z = \frac{1}{Y}$ og $W = \frac{1}{X}$ fås Z= a+b · W
6	Logarithmic-X model: Y = a + b*ln(X)	Sættes W = ln(X) fås Y= a+b · W
7	Multiplicative model: Y = a*X^b	$Y = a \cdot X^{b} \Leftrightarrow \ln(Y) = \ln(a) + b \cdot \ln(X)$ Sættes Z = ln(Y) og W = ln(X) fås Z= a+b · W
8	Square root-X model: Y = a + b*sqrt(X)	$Y = a + b \cdot \sqrt{X}$. Sættes $W = \sqrt{X}$ fås Y= a+b · W
9	Square root-Y model: Y = (a + b*X)^2	$Y = (a + b \cdot X)^2 \Leftrightarrow \sqrt{Y} = a + b \cdot X$ Sættes $Z = \sqrt{Y}$ fås Z= a+b · X
10	S-curve model: Y = exp(a + b/X)	$Y = e^{a + \frac{b}{X}} \Leftrightarrow \ln(Y) = a + \frac{b}{X}$ Sættes Z = ln(Y) og $W = \frac{1}{X}$ fås Z= a+b · W
11	Logistic model: Y = exp(a + b*X)/(l + exp(a + b*X))	$Y = \frac{e^{a+bX}}{1 + e^{a+bX}} \Leftrightarrow \ln\left(\frac{1}{Y} + 1\right) = a + bX$ Sættes $Z = \ln\left(\frac{1}{Y} + 1\right)$ fås Z= a+b · X
12	Log probit model: Y = normal(a + b*ln(X))	$Y = \Phi(a + b \cdot \ln(X)) \Leftrightarrow \Phi^{-1}(Y) = a + b \cdot \ln(X)$ Sættes $Z = \Phi^{-1}(Y)$ og W=ln(X) fås Z= a+b · W

Har man på forhånd en viden om, at en bestemt transformation skal anvendes, kan man uden større besvær foretage den pågældende transformation på dataene og så udføre regressionsanalysen på de transformerede data, på samme måde som vist i eksempel 15.5.
Dette illustreres ved følgende eksempel.

Eksempel 15.6. Valg mellem lineær og eksponentiel model

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_t 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 funktioner overvejes:
Model l (lineært henfald): $C = a + b \cdot t$
Mode12 (eksponentielt henfald): $C = a \cdot e^{b \cdot t}$

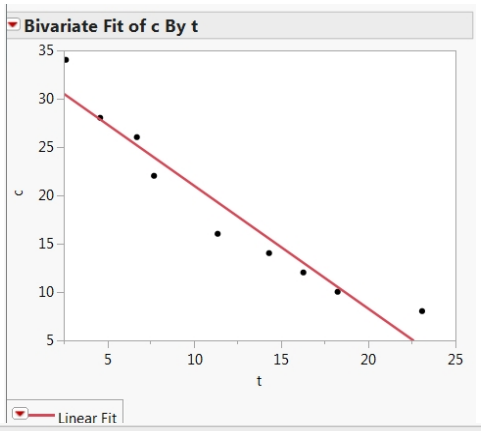
1) Vurder hvilken model der er bedst.
2) Opskriv regressionsligningen for den model du finder bedst.
3) Beregn ud fra den valgte model den værdi af C, for hvilken t = 12 minutter, og opskriv et 95% konfidensinterval for C.

Løsning:
Data indtastes

t	c
2,67	34
4,59	28
6,75	26
7,67	22
11,34	16
14,34	14
16,25	12
18,25	10
23,09	8

1) Analyze ▶ Fit Y by X ▶ markér "c" og tryk på Y Response ▶ markér "t" og tryk på X Factor ▶ OK
Der fremkommer et "scatterplot", hvor man i et koordinatsystem kan se punkterne afbildet.
Rød pil ▶ Fit line

Der fremkommer følgende figur og udskrift:



Linear Fit

c = 33,710679 - 1,2710444*t

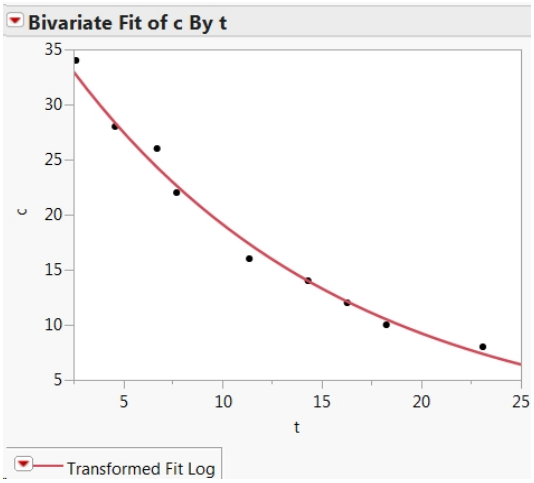
Summary of Fit

RSquare	0,929312
RSquare Adj	0,919214
Root Mean Square Error	2,559818
Mean of Response	18,88889
Observations (or Sum Wgts)	9

Det ses, at punkterne ikke fordeler sig jævnt om linien

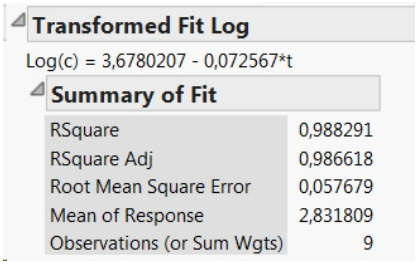
Forklaringsgraden 91.9% er høj

Vi gentager nu, idet vi nu vælger
Fit special ▶ Marker "Natural logarithm" for y ▶ OK



Vi ser, at punkterne fordeler sig tilfældigt omkring kurven,

og at RSquare =0.988 er høj, så
model2 (eksponentiel model) må være den bedste model.

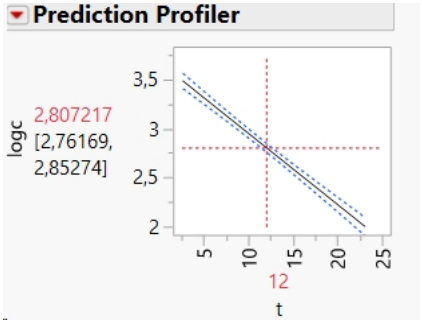


2) Af Log(c) = 3,6780207 - 0,072567*t fås (med regnemaskine)

$$c = e^{3.6780-0.07257\cdot t} = \underline{\underline{39.57 \cdot e^{-0.07257\cdot t}}}$$

3) Danner en ny kolonne med navnet logc ▶ cursor på navn, højre musetast, Formula ▶ trancental ▶ - vælg c ▶ ln ▶ ok

Analyze ▶ Fit model ▶ Marker logc og vælg y ▶ marker t og vælg add ▶ Emphasis, Minimal report ▶ Run
Rød pil ved "Response lnc" ▶ Factor profiling ▶ Profiler
Der fremkommer en tegning, t-værdi sættes til 12



Heraf ses, at for t = 12 er log(c) = 2.8072 $\Leftrightarrow c = e^{2.8072} = \underline{\underline{16.56}}$

95% konfidensinterval: $\left[e^{2.76169}; e^{2.8527}\right] = \underline{\underline{[15.83;17.33]}}$ ◆

15.6. ENKELT REGRESSIONSANALYSE MED FLERE y -VÆRDIER FOR HVER x - VÆRDI.

Hvis man selv fastlægger sine x - niveauer, er det ofte muligt for hver x -værdi, at foretage flere målinger af y - værdien. Vi siger kort at analysen er "med gentagelser". Sædvanligvis er det bedre at nøjes med en y-værdi for hver x-værdi, og så til gengæld have flere forskellige punkter, men i tilfælde, hvor man er i alvorlig tvivl om forudsætningen om varianshomogenitet gælder, eller hvor man vil teste om en lineær model gælder, kan det være relevant med eksempelvis to gentagelse for hver x-værdi.

Dette er således tilfældet i følgende eksempel:

Eksempel 15.7 Regressionsanalyse med gentagelser

Metalpladers overflader oxideres i en ovn ved 200^0 C. Med henblik på en undersøgelse af sammenhængen mellem det oxiderede lags tykkelse y (i ångstrøm) og tiden t (i minutter) foretog man følgende målinger:

Tiden t	20	30	40	60	70	90	100	120	150	180
Tykkelse y	4.2 4.9	7.4 6.9	8.8 8.2	13.6 12.0	13.1 12.4	14.9 16.8	20.0 21.2	23.1 25.2	27.5 25.1	32.9 32.4

Fordelen herved er, at man nu kan få et estimat for forsøgsfejlens spredning ("støjen"), som kan anvendes til at teste, om den lineære model kan accepteres, når man tager støjen i betragtning.

Endvidere kan man, hvis man finder det nødvendigt, teste om der er varianshomogenitet.
Alle andre test udføres på samme måde som beskrevet i forrige afsnit.

Forklaring af metode og formler

Test af model.

For hver x - værdi beregnes gennemsnittet af de dertil hørende y - værdier. Disse "gennemsnits-punkter" bør ligge tæt på linien hvis modellen er god. Hvis modellen er den rigtige, så er den eneste grund til at "gennemsnitspunkterne ikke ligger eksakt på linien, at der er støj.

Vi kan derfor beregne et estimat (kaldet $s^2_{\text{lack of fit}}$) for variansen af denne støj ud fra de afvigelser som gennemsnitspunkterne har. Hvorledes denne beregnes ses i notatet "Supplement til Anvendt statistik" som findes på min hjemmeside .

Da vi samtidig ud fra gentagelserne kan beregne et andet estimat for støjen (kaldet s^2_{error}), har vi mulighed for at teste de to varianser mod hinanden, ved en sædvanlig F - test

$$F_{\text{lack of fit}} = \frac{s^2_{\text{lack of fit}}}{s^2_{error}}$$

Får vi her en forkastelse, kan "gennemsnitspunkternes" afvigelser fra linien ikke forklares alene ved støjen, og vi må derfor forkaste modellen.

Test af varianshomogenitet.
Som tidligere nævnt, er analysen robust overfor afvigelser fra kravet om varianshomogenitet (konstant varians σ^2), hvis der er lige mange gentagelser (som i dette forsøg). Man vil derfor kun foretage en vurdering af dette krav, hvis man ud fra forsøgets natur mener, at varianserne kan tænkes at være voldsomt forskellige.
Testen kan foretages præcist som beskrevet under ensidet variansanalyse (se).

Eksempel 15.7. Regressionsanalyse (med gentagelser)
Givet følgende målinger

Tiden t	20	30	40	60	70	90	100	120	150	180
Tykkelse y	4.2 4.9	7.4 6.9	8.8 8.2	13.6 12.0	13.1 12.4	14.9 16.8	20.0 21.2	23.1 25.2	27.5 25.1	32.9 32.4

1) Foretag en testning af forudsætningen om varianshomogenitet.
2) Det formodes på forhånd, at der er en lineær sammenhæng mellem x og y.
   Undersøg ved en "lack of fit" test, om formodningen kan accepteres.
3) Bestem i bekræftende fald ligningen for den fundne regressionslinie.
4) Det påstås i litteraturen, at hældningskoefficienten β_1 er 0.15
   Test om dette på et signifikansniveau på 5% kan være sandt.
5) Angiv et 95% konfidensinterval for middelværdien af tykkelsen y, når t = 110 minutter.

Løsning
Data indtastes på sædvanlig måde:

t	y
20	4,2
20	4,9
30	7,4
30	6,9
40	8,8
40	8,2
60	13,6

osv.

1) Undersøgelse af varianshomogenitet.
   Test af nulhypotesen $H_0: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_{10}^2$
   Man gør som beskrevet under ensidet variansanalyse, dvs. vi vælger at gøre t til"nominal" (rød)
   Vælg Fit Model ▶ Fit Y by X ▶ markér "y" og tryk på Y Response ▶ markér "t" og tryk på X Factor ▶ OK
   Sæt cursor på overskrift, højre musetast ▶ vælg fra rullemenuen " "UnEqual Variances".

Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	0,0000	-1	0	0,0000
Brown-Forsythe	-1,9e+16	9	10	0,0000
Levene	-1,9e+16	9	10	0,0000
Bartlett	0,4990	9	.	0,8763

Warning: Small sample sizes. Use Caution.

Da vi kun har 2 gentagelser for hver t-værdi kan kun Bartletts test anvendes.
Da P - værdien=0.8763 > 0.05 accepteres H_0 , dvs. vi vil i det følgende antage, at kravet om varianshomogenitet er opfyldt.

2) H_0: Lineær model gælder $\Leftrightarrow$ H_0: (x_i, μ_i) ligger på en ret linie

Man sørger nu for, at t er "continous", og vælger nu forfra

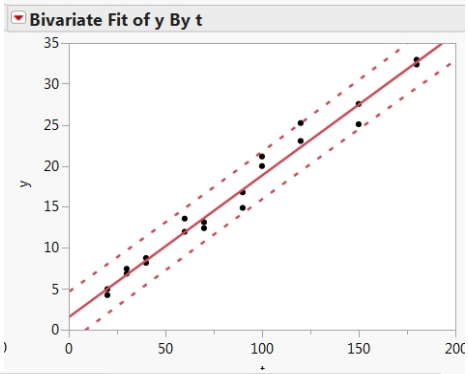
Analyze ▶ Fit Y by X ▶ markér "y" og tryk på Y Response ▶ markér "t" og tryk på X Factor ▶ OK

Der fremkommer et "scatterplot", hvor man i et koordinatsystem kan se punkterne afbildet.

Sæt cursor på overskrift, højre musetast ▶ vælg fra rullemenuen "Fit Line".

Tryk på den røde pil under tegningen ved "linear fit" ▶ vælg"Confid. Curves indv"

Der fremkommer følgende udskrift:



Af figuren ses, at "gennemsnitspunkterne" ligger tilfældigt omkring linien, og der næppe er outliers, da punkterne næsten alle falder indenfor "prediction linierne"

Linear Fit

y = 1,6541465 + 0,172975*t

Summary of Fit

RSquare	0,978579
RSquare Adj	0,977389
Root Mean Square Error	1,344501
Mean of Response	16,53
Observations (or Sum Wgts)	20

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	8	22,728289	2,84104	2,8961
Pure Error	10	9,810000	0,98100	Prob > F
Total Error	18	32,538289		0,0591
				Max RSq
				0,9935

Det ses, at forklaringsgraden RSquare er høj

Af udskriften for "Lack of fit" ses, at P - value er 0.0591.

På et signifikansniveau på 5%, ses, at H_0 må accepteres, dvs. vi kan antage, at indenfor måleområdet giver førstegradsmodellen en rimelig god beskrivelse af resultaterne.

3) Af udskriften ses, at regressionsligningen bliver $\hat{y} = 1.6542 + 0.1730 \cdot x$

4) $H_0: \beta_1 = 0.15$

Sæt cursor på et vilkårligt tal under " Parameter estimates" , højre musetast ▶ Columns ▶ vælg "lower" ▶ gentag men vælg "Upper"

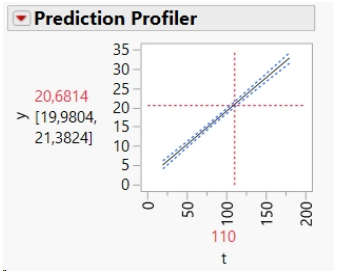
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	1,6541465	0,599582	2,76	0,0129*	0,3944709	2,9138221
t	0,172975	0,006032	28,68	<,0001*	0,160302	0,1856481

Heraf ses, at et 95% konfidensinterval for β ikke indeholder 0.15.
dvs. data giver ikke den i litteraturen angivne hældningskoefficient..

5) Find det til $t = 110$ svarende 95% konfidensinterval for tykkelsen y.
Analyze▶Fit model▶Marker y og vælg y▶marker t og vælg add▶Minimal report ▶Run
Rød pil ved "Response y▶ Factor Profiling▶Profiler▶på figur skriv 110

.



Vi får y's værdi for x = 110 = 20.68.
95% konfidensinterval [19.98 ; 21.38]

15.7 MULTIPEL REGRESSIONSANALYSE.

15.7.1 Indledning

Vi vil i dette afsnit behandle det tilfælde, hvor der indgår mere end 1 kvantitativ variabel. Vi vil begrænse os til at se på modeller, hvor de variable indgår lineært. Et eksempel herpå er modellen $Y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2$, hvor parametrene er α_0, β_1 og β_2.

15.7.2 Multipel regressionsanalyse med én *y* - værdi for hver *x* - værdi.

Man vurderer om modellen er acceptabel ved
1) at se på "forklaringsgraden" r^2
2) at se om der er outliers (se om "studentized residuals" alle numerisk er under 3, og kun få er over 2)

Vi vil illustrere metoden ved følgende eksempel.

Eksempel 15.8 (multipel regressionsanalyse uden gentagelser)

Det månedlige elektriske forbrug Y på en fabrik formodes at være afhængig af den gennemsnitlige udendørs temperatur x_1, antal arbejdsdage x_2 i måneden , den gennemsnitlige renhed x_3 af det fremstillede produkt og det antal tons x_4, der produceres i den pågældende måned. Det formodes, at Y er en lineær funktion af x_1, x_2 , x_3 og x_4 , dvs. på formen

$Y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$.

Følgende observationer fra det forløbne år foreligger

x_1	x_2	x_3	x_4	Y
- 4	22	91	100	836
-1	20	90	95	789
7	21	88	110	883
16	19	87	88	790
18	20	91	94	816
23	19	94	99	859
27	23	87	97	831
29	21	86	96	832
24	22	88	110	897
16	23	91	105	872
10	20	90	100	842
3	20	89	98	821

1) Vurder ud fra forklaringsgraden og "studentized residualer" om ovennævnte model er rimelig. Det antages i det følgende, at ovenstående model gælder.
2) Undersøg om modellen kan reduceres, dvs. kan nogle af koefficienterne antages at være 0.
3) Angiv regressionsligningen i den endelige model.
4) Angiv 95% konfidensintervaller for de regressionskoefficienter der indgår i ovenstående model
5) Angiv værdien og et 95% konfidensinterval for Y i punktet $(x_1, x_2, x_3, x_4) = (0,20,90,100)$

Løsning :
Data indtastes

	x1	x2	x3	x4	y
1	-4	22	91	100	836
2	-1	20	90	95	789
3	7	21	88	110	883
4	16	19	87	88	790
5	18	20	91	94	816
6	23	19	94	99	859
7	27	23	87	97	831
8	29	21	86	96	832
9	24	22	88	110	897
10	16	23	91	105	872
11	10	20	90	100	842
12	3	20	89	98	821

1) Vælg Analyze▶Fit model▶markér "y" og tryk på Y Response▶markér "x1, x2, x3, x4" og tryk ADD▶Emphasis: Minimal report▶Run
Der fremkommer bl.a. følgende udskrift

Response y

Summary of Fit

RSquare	0,965448
RSquare Adj	0,945705
Root Mean Square Error	7,90936
Mean of Response	839
Observations (or Sum Wgts)	12

Vælg rød pil ▶Save Columns▶Studentized residuals
I datatabel kan man nu yderligere finde følgende

x1	x2	x3	x4	y	Studentized Resid y
-4	22	91	100	836	1,7900396908
-1	20	90	95	789	-2,254767828
7	21	88	110	883	0,1362598723
16	19	87	88	790	1,4978025262
18	20	91	94	816	-0,633566393
23	19	94	99	859	0,5274844434
27	23	87	97	831	-0,653345676
29	21	86	96	832	0,1146976544
24	22	88	110	897	-0,330042097
16	23	91	105	872	0,2078113731
10	20	90	100	842	0,2523615824
3	20	89	98	821	-0,012553137

Da kun en enkelt værdi numerisk er større end 2 og ingen er over 3, antages, at der ikke er outliers
Da yderligere forklaringsgraden= 0.965 er tæt ved 1 vurderes modellen at være rimelig god.

2) Mulig reduktion af modellen
$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, H: Mindst en af regressionskoefficienterne er forskellig fra 0.

I samme udskrift som under "Summery of Fit" fandtes

Response y

Summary of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	12236,094	3059,02	48,8990
Error	7	437,906	62,56	Prob > F
C. Total	11	12674,000		<,0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	175,49949	113,8626	1,54	0,1671
x1	1,0266425	0,226819	4,53	0,0027*
x2	-0,793015	2,120284	-0,37	0,7195
x3	1,9113126	1,145911	1,67	0,1393
x4	4,9822626	0,44901	11,10	<,0001*

Af ovenstående udskrift ses for model, at P - Value <0.0001 < 0.05.
Heraf følger, at H_0 forkastes (stærkt), dvs. mindst en af regressionskoefficienterne er forskellig fra 0.
Vi ser nu regressionskoefficienterne
Den β størrelse, der har størst P-værdi er β_2.
H_0: $\beta_2 = 0$ accepteres, da P -værdien = 0.7195> 0.05.
x2-leddet bortkastes.
Bemærk, at man kun eliminerer én variabel ad gangen.

Vi eliminerer nu x_2 :
(slettes ved at trykke på "Remove")

Da *P*-værdien for x3 er 0.0849 > 0.05 eliminerer vi nu x3.

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
x1	1	1	1274,177	22,8216	0,0014*
x3	1	1	215,715	3,8636	0,0849
x4	1	1	10796,609	193,3764	<,0001*

Det er nu ikke muligt at reducere modellen mere.

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
x1	1	1	1083,972	14,7285	0,0040*
x4	1	1	11089,679	150,6814	<,0001*

3) For at kunne angive regressionsligningen betragtes følgende udskrift.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	335,65234	40,26796	8,34	<,0001*
x1	0,9016966	0,234953	3,84	0,0040*
x4	4,9401735	0,40245	12,28	<,0001*

Ligningen bliver $\hat{y} = 335.65 + 0.9017 \cdot x_1 + 4.9402 \cdot x_4$

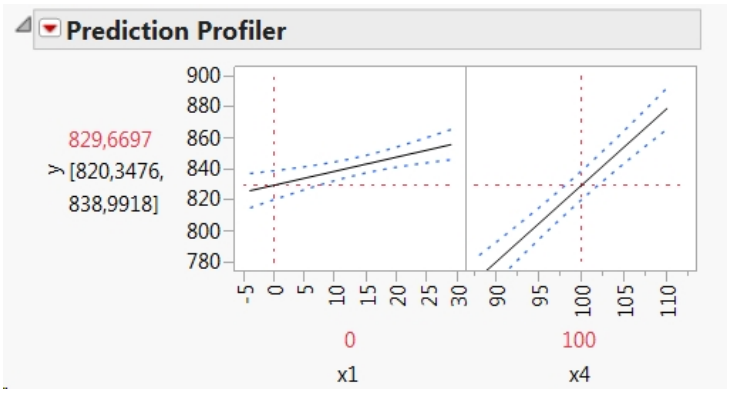
4) Cursor i tabellen ovenfor, højre musetast ▶Columns▶Upper 95% ▶lower 95%

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	335,65234	40,26796	8,34	<,0001*	244,55989	426,74479
x1	0,9016966	0,234953	3,84	0,0040*	0,370196	1,4331972
x4	4,9401735	0,40245	12,28	<,0001*	4,0297676	5,8505795

Konfidensintervallerne bliver $\beta_1 : [0.3702 ; 1.4332]$, $\beta_4 : [4.030 ; 5.851]$

5) Cursor på rød pil"Response Y" ▶vælg "Factor Profiling" ▶ Profiler▶ cursor på det røde tal forneden ved x1, skriv 0, Cursor på rødt tal ved x4, skriv 100



Vi har derfor, at $\hat{y} = 829.67$ og et 95% konfidensinterval er [820.35 ; 838.99]

15.8 POLYNOMIAL REGRESSIONSANALYSE

15.8.1 Indledning

Ved en polynomial regressionsanalyse er den statistiske model

$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots + \beta_p x^p$.

hvor den variable Y skal opfylde de sædvanlige regressionsforudsætninger.

Som det ses, er den i afsnit 15.2 betragtede enkelte regression et specialtilfælde. Den statistiske analyse da også meget beslægtet hermed.

Det man søger er altid den "enkleste" model der giver en tilstrækkelig god beskrivelse af Y indenfor det foreliggende variationsområde for x. Ud fra et statistisk synspunkt, vil man altid foretrække den model med de færreste parametre, da de på samme datamateriale giver en sikrere bestemmelse af parametrene. At andet lige vil man derfor foretrække de i afsnit 15.3 nævnte "transformerede" modeller som alle kun har 2 parametre fremfor eksempelvis et andengradspolynomium $Y = \beta_0 + \beta_1 x + \beta_2 x^2$ hvor man skal bestemme tre parametre β_0, β_1 og β_2.

Blandt polynomierne vil man naturligvis foretrække et af lavest grad.

Programmerne tegner let polynomier af 2. grad , 3 grad og 4 grad og regner forklaringsgraden ud.

Derved kan man få et indtryk af hvilket polynomium, af lavest grad, der giver en rimelig beskrivelse af data.

Imidlertid er man altid nødt til at teste modellen, og derfor er fremgangsmåden den, at polynomiet omskrives og man går over til multipel regression.

Eksempelvis omskrives polynomiet $T = a_0 + b_1 x + b_2 x^2 + b_3 x^3$

til $T = a_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$ ved at sætte $x_1 = x, x_2 = x^2, x_3 = x^3$

Vi vil i dette kapitel kun betragte tilfældet hvor der til hver x-værdi svarer en y-værdi.

Modificeret forklaringsgrad (adjusted)

Et 17-gradspolynomium vil gå eksakt gennem de 18 punkter, og $r^2 = 100\%$. Det er imidlertid klart, at en sådan model dels er alt for komplicerede til de fleste praktiske formål, dels følger kurven alle de tilfældige variationer, som vi netop ikke bør tage hensyn til.

Betragter man nu eksempelvis polynomiet $Y = \beta_0 + \beta_1 x + \beta_2 x^2$, beregner forklaringsgraden r^2 , derefter betragter et polynomium af tredie grad, og beregner igen forklaringsgraden , så er den steget, selv om trediegradsmodellen måske ikke er væsentlig bedre.

Det er altid muligt at øge r^2 ved at addere flere led til modellen.

For hvert led der tilføjes mistes der en frihedsgrad i "error", og hvis SAK for det nye led ikke giver et væsentligt bidrag kan det betyde, at den nye model er ringere end den gamle model. For at tage hensyn til dette, betragtes ofte et modificeret r^2 (R-squared (adjusted for d.f.[1]))

Når r^2 ikke stiger væsentligt, og R-squared (adjusted for d.f.) næsten er konstant, eller begynder at falde, er man tæt ved den "bedste" model.

---

[1] r^2 (adjusted) $= \dfrac{(n-1)\cdot r^2 + 1 - k}{n-k} = \dfrac{s_{total}^2 - s_0^2}{s_0^2}$, hvor k er antal parametre i modellen(incl konstantled).

Fremgangsmåde
Man starter med at tegne punkterne , eventuelt sammen med en lineær model af første grad $y = \alpha + \beta_1 x$ (se eksempel 15.4)
Er punkterne i et interval stigende og i et andet interval faldende så de eksempelvis kunne beskrive en parabel, så vil det være rimeligt at se på et polynomium af anden eller højere grad.
Der er i princippet to fremgangsmåder:
Forward:
Man ser på en model af anden grad $y = \alpha + \beta_1 x + \beta_2 x^2$ og sammenligner den med en af tredie grad $y = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$
Dette sker ved at man sammenligner de to "adjusted forklaringsgrader ",
samt teste nulhypotesen $H_0 : \beta_3 = 0$
Hvis man får en accept af H_0 og endvidere at "adjusted r^2 " for trediegradspolynomiet ikke er væsentlig højere end den af anden grad, så vil man vælge andengradspolynomiet.
Man vil dog så for andengradspolynomiet også se om der er outliers (studenticed residuals).
Hvis man omvendt får en forkastelse af $H_0 : \beta_3 = 0$ vil man sammenligne trediegradspolynomiet med et polynomium af fjerdegrad.

Backward
Her starter man med en model af fjerde grad $y = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4$
Man tester nulhypotesen $H_0 : \beta_4 = 0$
Får man en accept af nulhypotesen ser man nu på en model af trediegrad, og tester analogt $H_0 : \beta_3 = 0$
Endvidere sammenligner man de to "adjusted r^2 " for at se om der er sket en væsentlig ændring
Således fortsættes indtil man får en forkastelse af nulhypotesen, samt ser, at "adjusted r^2 " ikke har ændret sig væsentlig.

De to metoder kan godt give forskelligt resultat, da der jo er et element af personlig vurdering i konklusionen.
For det polynomium man ender med, ser man derfor på outliers, og residualplot for at være sikker på, at modellen er acceptabel..

15.8.2. Beregning af polynomial regressionsanalyse

Eksempel 15.9. Polynomial regressionsanalyse uden gentagelser.

Et forsøg udføres, for at finde hvordan størkningstiden T (i minutter) afhænger af antal gram x af et additiv. Man fik følgende forsøgsresultater:

x g/l	0	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8	8.5
T min.	740	710	610	650	470	540	440	420	400	450	440	480	530	470	420	480	450	490

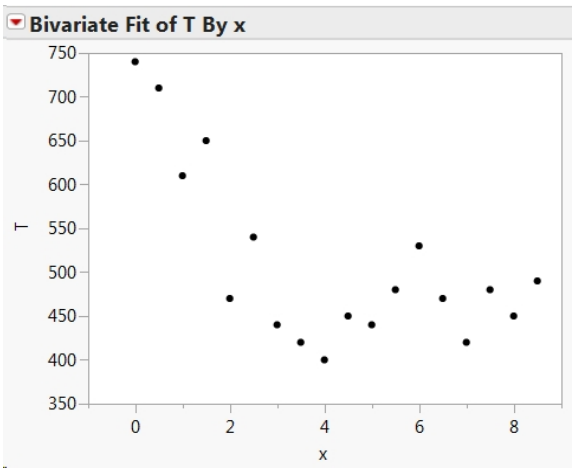
1) Vurder på basis af ovennævnte observationer, hvilket polynomium
$T = \alpha_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots + \beta_p x^p$ af lavest mulig grad p, der indenfor måleområdet [0 ; 8.5] giver en tilfredsstillende beskrivelse af T's variation.
2) Angiv regressionsligningen for den model, man i spørgsmål 1 har fundet frem til.
3) Beregn værdien af T for x = 6.2, og angiv et 95% konfidensinterval for T for x = 6.2.

Løsning:

1) Data indtastes

Analyze ▶ Fit Y by X ▶ markér "T" og tryk på Y Response ▶ markér "x" og tryk på X Factor ▶ OK

Der fremkommer et "scatterplot", hvor man i et koordinatsystem kan se punkterne afbildet.



Cursor på rød pil ved overskrift på tegning, højre musetast ▶ Fit Polynomial

Man kan nu vælge, hvilken grad polynomiet skal have.

Ud fra scatterplottet synes en andengradsmodel ikke at være en god model

Vi vælger dog først en andengradsmodel 2.quadratic

Der fremkommer blandt andet følgende udskrift:

Summary of Fit	
RSquare	0,787571
RSquare Adj	0,759247
Root Mean Square Error	49,711
Mean of Response	510,5556
Observations (or Sum Wgts)	18

Vi vælger nu en trediegradsmodel 3,cubic

Vi ser, at R-squared (adjusted) er steget fra 0.759 til 0.807.

Summary of Fit	
RSquare	0,841256
RSquare Adj	0,807239
Root Mean Square Error	44,48109
Mean of Response	510,5556
Observations (or Sum Wgts)	18

Vi vælger nu en fjerdegradsmodel 4,quartix

Vi ser, at RSquar Adj (adjusted) er faldet fra 0.807 til 0.799

Summary of Fit	
RSquare	0,846337
RSquare Adj	0,799056
Root Mean Square Error	45,41548
Mean of Response	510,5556
Observations (or Sum Wgts)	18

Heraf må sluttes, at fjerdegradsmodellen ikke har givet et væsentligt forbedret bidrag til forklaring af data.

For at lave tests, konfidensintervaller m.m. må man indføre 2 nye kolonner x2 = x^2 og x3=x^3 og gå over i multipel analyse.
Hertil benyttes formula (Cursor på kolonneoverskrift ▶ højre musetast ▶ formula) ▶vælg potensformel i oversigt

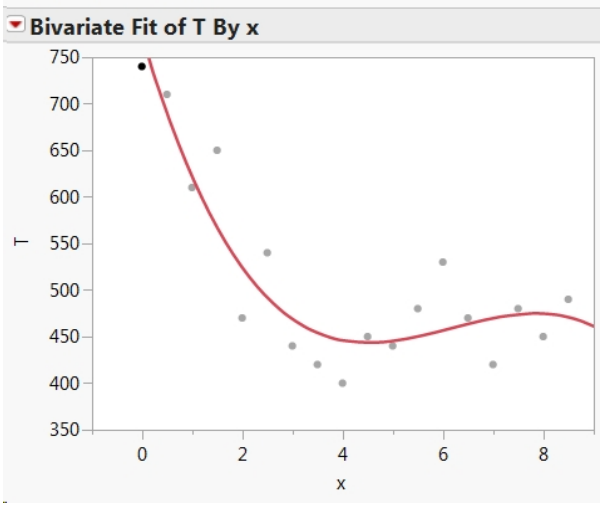
Vælg Analyze▶Fit model▶markér "T" og tryk på Y Response▶markér "x, x2, x3" (vælg minimal report) og tryk ADD ▶ Run
Man får bl.a.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	770,70175	34,52201	22,32	<,0001*
x	-179,2699	36,21048	-4,95	0,0002*
x2	31,336429	10,07039	3,11	0,0077*
x3	-1,692466	0,777816	-2,18	0,0472*

Da vi ser, at P-værdien for $\beta_3 = 0.047 < 0.05$ forkastes H_0: $\beta_3 = 0$, dvs.
Vi kan ikke bortkaste trediegradsleddet.
Heraf sluttes, at en trediegradsmodel må være det foreløbig bedste bud
Grafen for trediegradsmodellen bliver

Da punkterne ligger tilfældigt omkring kurven finder vi, at trediegradsmodellen er en acceptabel model



For at vurdere om der er outliers, vælges studentized residuals
Vælg rød pil(ved response) ▶Save Columns▶Studentized residuals
Resultat blev følgende tabel

x	T	x2	x3	Studentized Resid T
0	740	0	0	-1,094539224
0,5	710	0,25	0,125	0,5615462107
1	610	1	1	-0,273772112
1,5	650	2,25	3,375	2,0531565861
2	470	4	8	-1,337807591
2,5	540	6,25	15,62	1,19147889
3	440	9	27	-0,717878586
3,5	420	12,2	42,87	-0,838797886
4	400	16	64	-1,123753994
4,5	450	20,2	91,12	0,1366190753
5	440	25	125	-0,150579153
5,5	480	30,2	166,3	0,7109070448
6	530	36	216	1,7941894139
6,5	470	42,2	274,6	0,133409821
7	420	49	343	-1,250031893
7,5	480	56,2	421,8	0,1274797235
8	450	64	512	-0,672767712
8,5	490	72,2	614,1	0,6567235345

Den viser, at ingen numerisk er over 3, og kun en enkelt på 2.05 er numerisk over 2

Konklusion: Trediegradsmodellen er den bedste model

2) Regressionsligningen ses under parameter Estimates

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	770,70175	34,52201	22,32	<,0001*
x	-179,2699	36,21048	-4,95	0,0002*
x2	31,336429	10,07039	3,11	0,0077*
x3	-1,692466	0,777816	-2,18	0,0472*

$$\hat{T} = 770.70 - 179.27x + 31.336x^2 - 1.629x^3 .$$

3) Beregn værdien af T for x = 6.2, og angiv et 95% konfidensinterval for T for x = 6.2.

Skriv 6.2 nederst i datatabel ▶ cursor på overskrift på x2, højre musetast ▶

Nu kommer automatisk i kolonnerne 6.2^2 osv.

Vælg Analyze ▶ Fit model ▶ osv

Rød pil ved Response T ▶ Save Columns ▶ Predicted value

derefter det samme men nu vælges ▶ Mean Confidence Interval

Nu dannes i tabellen tre nye kolonner op, hvoraf man aflæser det ønskede.

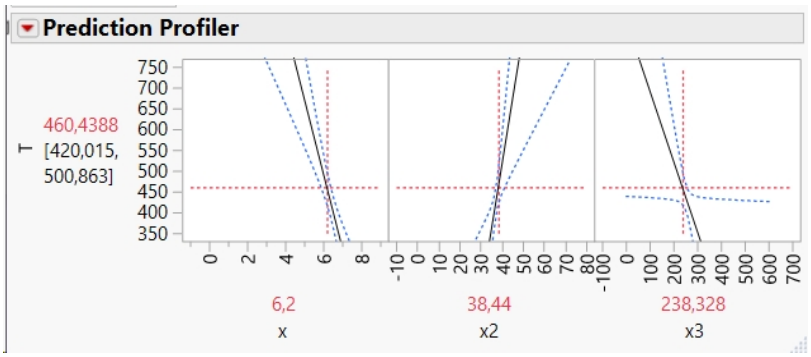
x	T	x2	x3	Studentized Resid T	Predicted T	Lower 95% Mean T	Upper 95% Mean T
0	740	0	0	-1,094539224	770,70175439	696,65941622	844,74409255
0,5	710	0,25	0,125	0,5615462107	688,68937049	638,92422323	738,45451774
1	610	1	1	-0,273772112	621,07585139	581,42093629	660,73076651
1,5	650	2,25	3,375	2,0531565861	566,59184727	527,7344981	605,44919643
2	470	4	8	-1,337807591	523,96800826	483,77394584	564,16207067
2,5	540	6,25	15,625	1,19147889	491,93498452	451,74092211	532,12904693
3	440	9	27	-0,717878586	469,22342621	430,77106308	507,67578935
3,5	420	12,25	42,875	-0,838797886	454,56398349	418,63727882	490,49068816
4	400	16	64	-1,123753994	446,6873065	412,60539012	480,76922288
4,5	450	20,25	91,125	0,1366190753	444,32404541	410,24212903	478,40596179
5	440	25	125	-0,150579153	446,20485036	410,27814569	482,13155503
5,5	480	30,25	166,37	0,7109070448	451,06037152	412,60800838	489,51273465
6	530	36	216	1,7941894139	457,62125903	417,42719662	497,81532144
6,5	470	42,25	274,62	0,133409821	464,61816305	424,42410064	504,81222547
7	420	49	343	-1,250031893	470,78173375	431,92438458	509,63908292
7,5	480	56,25	421,87	0,1274797235	474,84262126	435,18770615	514,49753636
8	450	64	512	-0,672767712	475,53147575	425,76632849	525,29662301
8,5	490	72,25	614,12	0,6567235345	471,57894737	397,5366092	545,62128554
6,2	•	38,44	238,32	•	460,43878225	420,01505899	500,86250551

Man får T = 460.44 og 95% konfidensinterval [420.0;500.9]

En anden mulighed er

Rød pil ved Resonse T ▶ Factor profiling ▶ Profiler

Der fremkommer to figurer. Tallene ændres i bunden



15.9. OVERSIGT OVER FREMGANGSMÅDE VED REGRESSIONSANALYSE

15.9.1 Lineær model:

Valg af model

Punkter + ret linie tegnes se eksempel 15.4

1) Punkter ligger tilfældigt om den rette linie
   a) Se på forklaringsgrad r^2
      a1) r^2"høj"(over ca. 70%), så ok
      a2) r^2 lav, så test om hældning er 0. se eksempel 15.5
          Hvis ja, så er y uafhængig af x (ingen sammenhæng)
          Hvis nej, så viser data en stigende / faldende tendens.
   b) Se på Outliers. Se evt. på "Studentized Residuals"
      b1) Ingen outliers, så ok
      b2) Outliers : Forsøgsomstændighederne ved dette forsøg undersøges.
2) Punkter ligger ikke tilfældigt om linie
   a) Se på forklaringsgrad r^2
      a1) r^2 meget høj (over 90%) :Model måske accepteres indenfor måleområdet.
      a2) r^2 ikke så høj, eller en sikrere model ønskes undersøgt:
          I: Punkter ligger "krummer monotont op eller ned":
          Se på modeller med 2 parametre (eksponentiel, potens, logaritme)
      Model skal opfylde krav om: Punkter ligger tilfældigt om kurve, r^2 høj, Ingen outliers
          II: Punkter ligger ikke monotont
          Se på polynomier.
          Model vælges af lavest grad. Fit Model, quadratic, beregne R-squaret(adjusted)
          evt. teste om koefficient til højestegradsled $\neq 0$

Test

1) Ret linie $y = \beta_0 + \beta_1 x$
   a) H_0: Y er uafhængig af $x \Leftrightarrow H_0$: Regressionslinien er vandret $\Leftrightarrow H_0: \beta_1 = 0$.
   b) Konfidensinterval for β_1:
   c) Konfidensinterval for y svarende til x =t:
   d) Prædistinationsinterval for 1 ny observation
2) Andre modeller i 2 parametre:
   Ligningen transformeres til et lineært udtryk se eksempel 15.6
   Indsætter nye lister med de transformerede variable
   Derefter som under punkt 1)
3) Polynomier af anden eller højere grad. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots$
   Ligningen omskrives til multipel lineær model $y = \beta_0 + \beta_1 x + \beta_2 x_2 + \ldots$ sæt $x_2 = x^2$, $x_3 = x^3$ osv
   Se på forklaringsgrad (bør være over 0.7)
   Undersøge om der er Outliers (ingen numerisk over 3)

Test

1) Reduktion af model:
   Hvis højestegradsleddet har en største P-værdi >5%
   Fjerner leddet, Gentag. Der elimineres én variabel ad gangen
2) Koefficienter : Findes i blist (se evt. under multipel regression)

15.9.2 Multipel regression

Se eksempel 15.8

Model acceptabel

Se på forklaringsgrad (bør være over 0.7)

Undersøge om der er Outliers (ingen numerisk over 3)

Test

1) Reduktion af model:
   Find det led hvis koefficient har den største P-værdi >5% Fjerner leddet
   Gentag med næste P-værdi Der elimineres én variabel ad gangen
2) Regressionskoeficienter : Findes i blist
3) 95% konfidensinterval for β_1:
4) Man finder 95% konfidensintervallet for y for given x-værdi

OPGAVER

Opgave 15.1

Nedenstående tabel angiver sammenhørende værdier af den "radiale" afbøjning X (i milliradianer) og den totale energiflux Y (i kilowatt) på et solvarmeanlæg.

x	16.66	16.46	17.66	17.50	16.40	16.28	16.06	15.93	16.60	16.41	16.17	15.92	16.04	16.19	16.71
y	271.8	264.0	238.8	230.7	251.6	257.9	263.9	266.5	229.1	239.3	258.0	257.6	267.3	267.0	263.8
x	16.62	17.37	18.12	18.53	15.54	15.70	16.45	17.62	18.12	19.05	16.51	16.02	15.89	15.83	
y	259.6	240.4	227.2	196.0	278.7	272.3	267.4	254.5	224.7	181.5	227.5	253.6	263.0	265.8	

1) Tegn en figur, hvor punkterne og en relevant ret linie er indlagt
2) Find r^2 og anvend denne, ovenstående figur samt om der er outliers, til vurdering af om en lineær model kan antages at gælde.

I det følgende antages, at forudsætningerne for en regressionsanalyse er opfyldt, og at den lineære model gælder.

3) Opskriv regressionsligningen.
4) Angiv et estimat og et 95% konfidensinterval for hældningen β_1.
5) Beregn er 95% konfidensinterval for middelfluxen y i det tilfælde, hvor den radiale afbøjning x er 16.5 milliradianer.

Opgave 15.2

Man ønskede på et universitet at undersøge om der var en sammenhæng mellem de point de studerende fik ved en indledende prøve i matematik, og de point de fik ved den afsluttende prøve i matematik. Resultaterne var

Student	1	2	3	4	5	6	7	8	9	10
Indledende prøve x	39	43	21	64	57	47	28	75	34	52
Afsluttende prøve y	65	78	52	82	92	89	73	98	56	75

1) Giv en vurdering af om en ret linie m: y = a + bx på rimelig måde kan beskrive denne sammenhæng. Undersøg specielt om der er outliers.
2) Idet det antages at ovenstående linie er en god model skal man angive en ligning for regressionslinien m.

Det antages i det følgende at forudsætningerne for en regressionsanalyse er opfyldt.

3) Test om y er uafhængig af x (signifikansniveau α=0.01)
4) Find er 95% konfidensinterval for hældningskoefficienten β_1 .
5) En student har opnået 50 point ved den indledende prøve. Forudsig indenfor hvilket interval denne student pointtal vil ligge ved den afsluttende prøve. (signifikansniveau α=0.05)
6) Angiv et 95% konfidensinterval for middelværdien af det pointtal som studenter opnår ved den afsluttende prøve, når de alle ved den indledende prøve har opnået 50 point.

Opgave 15.3

Ved en kemisk proces vides reaktionshastigheden v at afhænge af mængden x af et bestemt additiv, som virker som katalysator.

Ved et fuldstændigt randomiseret forsøg fandtes følgende observationer:

Tilsat mængde additiv x	0.001	0.2	0.5	1.0	1.5	2.0
Reaktionshastighed v	1	17	29	41	50	58

1) Foretag en vurdering, ud fra forklaringsgrad og graf, hvilken af de fire standardmodeller (lineær, potens, eksponential, logaritmisk), der bedst approksimerer data.
2) Angiv regressionsligningen for den valgte model
3) Find ved en additivtilsætning på 1.75, middelværdien af reaktionshastigheden og angiv et 95% konfidensinterval herfor.

Opgave 15.4

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.

a) Giv en vurdering af, om modellen y = a + bx (antal ulykker aftager lineært med alderen) på rimelig måde kan beskrive denne sammenhæng.
b) En trafikekspert mener, at modellen $y = \alpha \cdot e^{bx}$ (antal ulykker aftager eksponentielt med alderen) giver en bedre beskrivelse af modellen. Har vedkommende ret?
c) Bestem ligningen for den model du finder bedst.
d) Angiv ud fra ovennævnte ligning det forventede antal fart-relaterede ulykker som 50 - årige i middel vil forårsage i den givne periode.
e) Angiv et 95% konfidensinterval for middelværdien af ulykker for 50-årige.

Opgave 15.5

I et organisk-kemisk laboratorium undersøgte man forskellige reaktionskinetiske processer. Ud fra teoretiske overvejelser har man fundet frem til, at "middeludbyttet" (angivet i %'-enheder) af en bestemt kemisk forbindelse for $t > 5$ er approksimativt bestemt ved et udtryk af formen

$y = 100 - \alpha_0 \cdot \beta_1^t$, hvor t angiver reaktionstiden og y procesudbyttet.

For at efterprøve rigtigheden af de teoretiske overvejelser udførte man et forsøg med følgende resultater:

t	6.5	8.2	11.1	13.6	16.4	18.5	20.7	23.0	25.8	28.5	33.3
y	39.5	64.7	65.6	72.9	88.0	92.7	92.5	95.9	96.3	98.3	99.2

1) Omskriv ovennævnte udtryk for modellen således, at regressionsmodellen kan gøres lineær i parametrene ved en logaritmisk transformation.
2) Foretag den logaritmiske transformation og vurder såvel grafisk som ud fra forklaringsgraden om den formodede model (1) kan accepteres.
3) Foretag, idet det forudsættes, at modellen (1) gælder, en estimation af parametrene α_0 og β_1.
4) Opstil et 95% - konfidensinterval for middelværdien af udbyttet y svarende til t = 20.

Opgave 15.6

Ved en standardisering af et bestemt hormonpræparat behandler man et mindre antal mus med doser af forskellig størrelse og registrerer i hvert tilfælde tiden t, indtil musen dør. Fra tidligere undersøgelser ved man, at t er normalfordelt med konstant varians og med en middelværdi, som er en lineær funktion af logaritmen til dosis, dvs. $t = a + b \cdot \ln(\text{dosis})$.

Til brug for standardiseringen af et produktionsparti af præparatet blev foretaget 5 delforsøg, som gav følgende resultater:

dosis (antal enheder)	1585	2239	2884	5248	6918
t (timer)	8.70	6.20	8.22	2.94	3.88

1) Angiv et estimat for regressionslinien, hvor t er en funktion af 1ogaritmen ti1 dosis.
2) Opsti1 et 95% - konfidensinterval for koefficienten til logaritmen til dosis.
3) Opsti1 et 95% - konfidensinterval for midde1værdien af t for en dosis på 6300 enheder.

Opgave 15.7

Man har erfaring for, at jerns viskositet Y under smeltning afhænger af jernets siliciumindhold x. Man besluttede sig ti1 at foretage et forsøg med henblik på at undersøge denne sammenhæng nærmere. Ved forsøget foretoges 3 viskositetsmålinger for hver af 5 forskel1ige værdier af siliciumindholdet. Forsøgsresu1taterne var:

	x				
	125	150	175	200	225
Y	47.5 55.0 40	60.0 55.0 50.0	65.0 67.5 70.0	72.5 75.0 77.0	77.5 85.0 75.0

1) Kontroller om forudsætningerne om varianshomogenitet er rimelig opfyldt.

Det antages i det følgende at forudsætningerne er rimelig opfyldt

2) Test om der er en lineær sammenhæng mellem jerns viskositet og siliciumindholdet, og angiv i bekræftende fald ligningen for den empiriske regressionslinie.

Det antages i det følgende, at der er en lineær sammenhæng mellem x og y.

3) Foretag en testning af om regressionslinien er vandret.
4) Angiv et 95% konfidensinterval for hældningskoefficienten
5) Angiv et 95% konfidensinterval for middelværdien af middelviskositeten y, når $x = 160$.

Opgave 15.8

Koncentrationsbestemmelse af stoffet aprindin kan foretages ved hjælp af en gaskromatograf. Ved denne metode indsprøjtes en del af prøven indeholdende aprindin i gaskromatografen, og den såkaldte tophøjde bestemmes. Såfremt de laboratorietekniske procedurer er korrekt udført, skal tophøjden, bortset fra tilfældige udsving, være proportional med koncentrationen i prøven. I et eksperiment fremstillede man 12 prøver med kendte koncentrationer af aprindin og målte tophøjderne. Resultaterne fremgår af nedenstående tabel.

	Koncentration x ($\mu g / ml$)					
	0.5	1	2	3	4	5
Tophøjde Y	46 55	120 90	232 241	310 318	440 435	550 539

1) Bestem den lineære regressionslinie for Y på x
2) Test, om en sådan lineær regression kan beskrive data.
3) Test, om tophøjden y kan antages at være proportional med koncentrationen x, dvs. $y = a \cdot x$.

Opgave 15.9

Den tid (y) det tager inden en bestemt maskinkomponent svigter kan tænkes at afhænge af den spænding (x_1) , den temperatur (x_2) som komponenten udsættes for under kørslen, samt motorens omdrejningshastighed pr. minut (x_3)

Det forløbne år har givet de data, som er vist i følgende tabel:

(x_1, x_2, x_3)	(110,60,750)	(110,82,850)	(110,60,1000)	(110,82,1100)	(120,60,750)
y	2145	2155	2220	2225	2360

(x_1, x_2, x_3)	(120,82,850)	(120,60,1000)	(130,82,1100)	(115,66,840)	(115,66,880)
y	2266	2334	2340	2212	2180

Det forudsættes, at regressionsforudsætningerne er opfyldt.

1) Vurder ud fra forklaringsgraden , og ud fra en vurdering af om der er outliers, om en lineær model i de tre variable ,dvs. af formen $Y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ er rimelig.

Det antages i det følgende, at ovenstående model gælder.

2) Undersøg om modellen kan reduceres, dvs. om nogle af koefficienterne kan antages at være 0.
3) Angiv regressionsligningen i den endelige model.
4) Bestem et estimat for Y i tilfældet $x_1 = 125$, $x_2 = 70$ og $x_3 = 900$, og angiv et 95% konfidensinterval for denne værdi.

Opgave 15.10

Ved en given produktion ønskes undersøgt, hvorledes mængden Y af et uønsket biprodukt afhang af mængderne x_1, x_2 og x_3 af tre tilsætningsstoffer. Følgende forsøg blev foretaget (kodede tal):

(x_1, x_2, x_3)	(1,1,1)	(2,9,4)	(3,3,9)	(4,7,5)	(5,5 7)	(6,3,3)	(7,6,2)	(8,9,6)
Y	30	85	55	75	76	56	85	106

Det forudsættes, at regressionsforudsætningerne er opfyldt.

1) Vurder ud fra forklaringsgraden, og ud fra en vurdering af om der er outliers, om en lineær model i de tre variable ,dvs. af formen $Y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ er rimelig.

Det antages i det følgende, at ovenstående model gælder.

2) Undersøg om modellen kan reduceres, dvs. om nogle af koefficienterne kan antages at være 0.
3) Angiv regressionsligningen i den endelige model.
4) Bestem et estimat for Y i tilfældet $x_1 = 4$, $x_2 = 5$ og $x_3 = 6$, og angiv et 95% konfidensinterval for denne værdi.

Opgave 15.11

Det formodes, at den producerede mængde Y af et stof ved en given produktion er en lineær funktion af de anvendte mængder x_1 , x_2, og x_3 af tre råvarer.

Følgende ikke særligt systematiske observationer foreligger:

x_2		1			2			4			5		
x_3		1	3	5	1	3	5	1	3	5	1	3	5
x_1	1	6.37				6.70							
	2			8.02							7.54		
	3		8.70				9.08						
	5				10.30				10.40				
	10							15.34					16.12

1) Vurder på basis af disse observationer, om en lineær model i x_1 , x_2 og x_3 er rimelig.
2) Foretag så vidt mulig en reduktion af modellen, og angiv til sidst regressionsligningen for den endelige model.
3) Beregn et 95% konfidensinterval for regressionskoefficienterne i den endelige model.
4) Beregn et 95% konfidensinterval for middelværdien af Y hvis $x_1 = 3$, $x_2 = 4$ og $x_3 = 2$.

Opgave 15.12

En fabrik fremstiller salpetersyre ved oxidering af ammoniak med luft. I løbet af processen ledes kvælstofoxider under afkøling ind i en absorptionskolonne, idet absorptionen i gennemstrømmende salpetersyre afhænger af kølevandstemperaturen x_1(°C), lufttemperaturen x_2(°C) og salpetersyre-koncentrationen x_3

Man ønsker at teste, om sammenhængen mellem mængden Y af ikke-absorberede kvælstofoxider i et givet tidsrum og x_1, x_2 og x_3 (aproksimativt) var lineær, og ønskede i bekræftende fald at estimere denne sammenhæng.

Følgende observationer af Y (kodede tal) fandtes:

				x_2			
				10	20	30	40
x_3	-5	x_1	5	32	50	62	88
			10	48	60	73	96
			15	54	70	90	102
	5	x_1	5	32	49	64	86
			10	50	64	82	92
			15	57	74	88	108

1) Vurder på basis af disse observationer, om en lineær model i x_1 , x_2 og x_3 er rimelig.
2) Undersøg, om modellen kan reduceres, dvs. om nogle af regressionskoefficienterne kunne være 0.
3) Opskriv regresionsligningen.
4) Angiv et 95% konfidensinterval for β_1.
5 Angiv et estimat for Y i tilfældet $x_1 = 8$, $x_2 = 20$ og $x_3 = 4$, og angiv et 95% konfidensinterval for denne værdi.

Opgave 15.13

Følgende sammenhørende data er 25 målinger mellem den jævnstrøm (y) en vindmølle udvikler og vindhastigheden (x).

x	5.00	6.00	3.40	2.70	10.0	9.70	9.55	3.05	8.15	6.20	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) Vurder på basis af ovennævnte observationer hvilket polynomium $y = \alpha_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots + \beta_p x^p$ af lavest mulig grad p, der indenfor måleområdet [0 ; 10] giver en tilfredsstillende beskrivelse af Y's variation.
2) Angiv regressionsligningen for den model, man i spørgsmål 1 har fundet frem til.
3) Skitser grafen indenfor intervallet [0;10], og beregn værdien af y svarende til en vindhastighed på x = 7, og et 95% konfidensinterval for x = 7

Opgave 15.14

Ved et fuldstændigt randomiseret forsøg foretoges følgende observationer mellem den ikke-statistiske variabel x og den statistiske variabel Y:

x	10	20	30	40	50	60	70	80
Y	72.5	78.3	79.6	78.9	76.9	76.6	68.9	66.4

1) Bestem ved en polynomial regressionsanalyse det polynomium i x af lavest mulig grad, der giver en tilfredsstillende beskrivelse af Y's variation.
   Betragt såvel forklaringsgrader som residualer
2) Opskriv regressionsligningen for den i spørgsmål 1 fundne model
3) Find den værdi x_m som giver den største y - værdi . Angiv endvidere den til x_m svarende estimerede middelværdi $\hat{Y}_m$, og et 95% konfidensinterval.

Opgave 15.15

Ved nogle forsøg med målinger af det tryk, som udgår fra jetmotorer, måltes for udvalgte værdier af ændringen i udstødningsdysens vinkel x værdier af ændringen i trykket Y. Resultaterne var:

x	4	4.5	5	6.5	7	7.3	7.5
Y(i%)	24.5	26.3	27.2	67.3	77.8	80.6	83.4

1. Bestem ved en polynomial regressionsanalyse det polynomium af lavest grad, der giver en tilfredsstillende beskrivelse af Y's variation.
2. Benyt den fundne model til at bestem middelværdien af Y, når $x = 6$ og beregn det dertil svarende 95% konfidensinterval .

FACITLISTE

KAPITEL 2
2.1 a) 65 0.4 b) 0.004

KAPITEL 3
3.1 (1) 0.35383 0.17966 0.06546 (2) 20.0243
3.2 (1) 69.15% (2) 10.88% (3) 112.2 (4) 117.3 6.535
3.3 (1) 86.64% (2) 0.00765 (3) [2.48 ; 2.52]
3.4 (1) 5.91% (2) 27.64% (3) [783.51; 816.49]
3.5 (1) 9.5 1.265 (2) 12.44 (3) 2.41%
3.6 (1) 92.75%
3.7 (1) 97.72% (2) 25.45
3.8 (1) 65 0.4 (2) 77.34%

KAPITEL 4
4.1 [74.036; 74.037]
4.2 (1) 12.13 0.6783 (2) [11.64 ; 12.62] (3) [10.52 ; 13.74]
4.3 (1) [82.50 ; 83.90] (2) ja
4.4 (1) 2259.92 35.569 (2) [2237.3 ; 2282.5] (3) [2178.4 ; 2341.4] (4) 34
4.5 (1) 19.13 (2) [14.0 ; 30.2] (3) 750.2 (4) [739.6 ; 760.8]
4.6 (1) 7.83 0.363 (2) [7.45 ; 8.22] (3) 51
4.7 (a) 4.256 [4.23 ; 4.29] (b) 66 (c) [4.16 ; 4.35] (d) 0,014 [0.0285 ; 0.0756]
4.8 (1) [96.36 ; 111.14] (2) [2.638; 17.322]

KAPITEL 5
5.1 (1) nej P-værdi = 2.67%
5.2 (1) ja P-værdi = 3.6% (2) 58 [55.82; 60.18]
5.3 (1) ja P-værdi = 0.44% (2) 84.47 [81.26 ; 87.67] (4) 25.72%
5.4 (a) nej P-værdi = 12.14% (b) -
5.5 (a) ja P-værdi = 0.157% (b) [24.07 ; 35.56]
5.6 nej P-værdi = 12.89% (evt. 6.44%)
5.7 ja, P-værdi < 0.0001
5.8 (1) 35 (2) ja P-værdi = 0.16% (3) nej [2.59 ; 2.67]
5.9 (1) 35 (2) ja P-værdi = $9.7 \cdot 10^{-15}$ (3) ja

KAPITEL 6
6.1 0.5 0.8 0.2 0.7
6.2 (1) 0.9134 (2) 0.9678
6.3 (1) 8.75% (2) 38.75% (3) 41.25% (4)11.25%
6.4 (1) 6.4% (2) 78.4% (3) 7.2%
6.5 $1.283 \cdot 10^{12}$
6.6 (a) - (b) 736
6.7 (a) 6 (b) 24

6.8 (a) 100 /b) 2400
6.9 60
6.10 (a) 30.24% (b) 0.24% (c) 99.76% (d) 4.04% (e) 44.04% (f) 21.44%
6.11 (1) 27.1% 36.0% 9.756% (2) 53.34% (3) 49.20%
6.12 3^{40}
6.13 (1) 31 (2) 9
6.14 30.24%
6.15 $9 \cdot 10^{7}$

KAPITEL 7
7.1 (1) 41.3% 2)- 3) 0.6
7.2 (A) 0.018% (B) 1.29% (C) 38.24%
7.3 a) 41.56% b) 21.94%
7.4 (1) ja (2) 49.74%
7.5 (1) 17.68% (2) 59.28%
7.6 2.58%
7.7 5.83%
7.8 5.69%
7.9 95.47%
7.10 40.33%
7.11 12.85%
7.12 44.16%
7.13 (1) 7.93% (2) 11.82%
7.14 (1) (a) 60.63% (b) 6.15% (2) 4.44%
7.15 nej, P-værdi =0.0079%
7.16 (1) 0.884% (2) nej
7.17 (1) ja, P-værdi = 2.53% (2) 0.12 (3) approx:[0.06 ; 0.19]
7.18 ja p = 0.43%
7.19 (1) 0.108 (2) [0.089 ; 0.128]
7.20 [0.799 ; 0.846]
7.21 1522
7.22 (1) 30.1% (2) 87.9% (3) 4
7.23 50.37%
7.24 (1) 15 (2) 81.9%
7.25 75.34%
7.26 6.56%
7.27 93.2%
7.28 44.57%
7.29 (1) ja P-værdi = 4.19% (2) 103.3 (4) [88.2 ; 118.8]
7.30 (1) 4.68 (2) [4.47 ; 4.89] (3) 69.44
7.31 (1a) 11 (1b) [5 ; 18] (2a) 1.1 (2b) [0.5 ; 1.8] eller [0.45 ; 1.75]

KAPITEL 8

8.1 (1) 2.90 (2) 14.6% (3) 16.98
8.2 (1) 77.88% (2) 10.45% (3) 77.88% (4) 9.48%
8.3 1
8.4 (1) 79.8% (2) 99.33% (3) 22.8 (4) 14
8.5 (1) 0.539% (2) 0.119%
8.6 7.31%
8.7 0.188%

KAPITEL 9

9.1 (1) NKG_R =0, $ØKG_R$ = 10.14, $NKG_{\bar{x}}$ = 451.4, $ØKG_{\bar{x}}$ = 457.8 (2) 1.03 %.
9.2 NKG_s =0, $ØKG_s$ = 1.410, $NKG_{\bar{x}}$ = 11.30, $ØKG_{\bar{x}}$ = 13.44
9.3 (1) - (2) - (3) $\bar{x}$ =2929.1 $\tilde{\sigma}$ =131.7
9.4 (1) NKG_R =0, $ØKG_R$ = 1.01, $NKG_{\bar{x}}$ = 85.26, $ØKG_{\bar{x}}$ = 86.42 (2) alarm 5 , 6 , 3
9.5 (1) 1.33 (2) 0,876% (3) 18,5 0.0066%
9.6 a) NKG=11.32, ØKG = 13.51 b) 1.068
9.7 NKG = 0.82, $ØKG$ = 19.85
9.8 ØKG = 16.09 NKG = 0
9.9 a) NKG = 0 ØKG = 15.16 7.14 b) 1.2 0.0167%
9.10 a) - b) alarm 3 , 2, 1
9.11 a) 1.179 b) -

KAPITEL 10

10.1 (1) - (2) 82.6% (3) 81.43%
10.2 (1) NKG=0, ØKG = 3.47 (2) c=4, n = 300 (3) 5.24 (4a) 0.70% (4b) 60.67%
10.3 (1) n=230, c=11 (2) ca. 78.7% (3) 78.7%

KAPITEL 12

12.1 *P* - værdi = 0.10
12.2 (1) *P* - værdi = 0.001 (2) [0.69 ; 3.13]
12.3 *P* - værdi = 0.0214
12.4 *P* - værdi =0.0472 , [0.015 ; 1.87]
12.5 a) *P* - værdi = 0.0001, b) 0.2258 C) [11.93 ; 14.96] c) 0.2258
12.6 (1) 87 (2) 87 (3) - (4) -
12.7 36
12.8 *P* - værdi = 0.589
12.9 *P* - værdi = 0.1398
12.10 *P* - værdi = 0.022
12.11 (1) *P* - værdi =0,024 (2) *P* - værdi =0.00334
12.12 *P* - værdi =0.233
12.13 (1) *P* - værdi = 0.2563 (2) *P* - værdi =0.0177
12.14 (1) *P* - værdi = 0.0134 (2) *P* - værdi = 0.0001
12.15 *P* - værdi = 0.00028

KAPITEL 13

13.1 a) - b) P - værdi = 0.0004 c) B2, 3181.5 [3055.4 ; 3307.1]
13.2 P - værdi = 0.0309
13.3 (1) P - værdi = 0.0064 (2) S2, S4
13.4 (1) P - værdi = 0.0006, R forskellig fra de øvrige, ja
13.5 (1) P - værdi = 0.0002 (2) T må foretrækkes.
13.6 (1) - (2) -
13.7 P - værdi = 0.0696
13.8 P - værdi =0.027

KAPITEL 14

14.1 (1) nej, P - værdi = 0.0108, (2a) K1 (2b) frit valg (2c) $A_3 K_2$, $A_2 K_1$, $A_1 K_2$,$A_3 K_1$
14.2 (1) ja (2) glas 2 og enten fosfor 1 eller 3.
14.3 (1) - 30.84% (2) limtyper (3)lim II eller lim III II: [23.76 ; 28.11],III: [22.69 ; 27.04]
14.4 (1) Kun syrer har virkning, Svovlsyre størst. (2) 26.45, [25.17 ; 27.73]
14.5 300^0 ,tryk 60
14.6 (1) Nej, P - værdi = 0.046 (2) ja, P - værdi = 0.078
14.7 Begge, Vitaminbehandling 1 eller 3.
14.8 (1) Nej, P-værdi = 0.0092 (2) råmateriale 4. (3) kun uger
14.9 Begge har signifikant virkning, vælge vævemetode 2, og ikke matrialtype 3.

KAPITEL 15

15.1 (1) - (2) 0.7205 (3) y= 607.1-21.4025 x (4) -21.4 [-26.66 ; -16.14] (5) [249.1 ; 258.8]
15.2 (1) 70.5% (2) y = 40.78 + 0.766 x (3) P - værdi = 0.0024 (4) [0.362 ; 1.169]
(5) [72.51 ; 85.61] (6) [57.95 ; 100.17]
15.3 (1) r^2 = 0.99 potens (2) $y = 40.673\ x^{0.53597}$ (3) 54.92 [53.36 ; 56.49]
15.4 (a) r^2 = 0.9606 (b) r^2 =0.9947 , Punkter ligger mere tilfældigt om kurve, Ja
(c) $y = 57.866\ 0.9707^x$ (d) 13.08 (e) [12.66 ; 13.52]
15.5 (1) - (2) r^2 = 0.9805 (3) $\widetilde{\alpha}_0 = 171.8$ $\widetilde{\beta}_1 = 0.8534$ (4) 92.786 [91.7 ; 93.7]
15.6 (1) $t = 35.435 - 3.6393 \cdot \ln(\text{dosis})$ (2) [-7.55 ;-0.2726] (3) [0.265 ; 6.929]
15.7 (1) - (2) P - værdi = 0.9746 (3) P(lack of fit) = 0.5453, $y = 6.683 + 0.3327\ x$
(4) P - værdi = $2.9 \cdot 10^{-7}$ (5) [25.77 ; 40.76] (6) [56.9 ; 62.69]
15.8 (1) Y = -0.1438 + 108.959 x (2) P - værdi = 0.144 (3) ja
15.9 (1) r^2 = 0.723 (2) - (3) $Y = 1131.8 + 9.5846 \cdot x_1$ (3) 2329.96 [2267 ; 2393]
15.10 (1) r^2 =0.9823 ja (2) ja (3) $Y = 21.625 + 3.4615x_1 + 6.2881x_2$ (4) 66.91, [62.74 ; 71.09]
15.11 (1) - (2) $y = 5.199 + 0.9964x_1 + 0.1689x_3$ (3)[0.964 ; 1.029], [0.107;0.231]
(4) 8.526, [8.406 ;8.646]
15.12 (1) - (2) - (3) $y = 5.9167 + 2.251x_1 + 1.648x_2$ (4) [1.92; 2.57] (5) 56.88, [55.33; 58.44]
15.13 (1) - (2) $Y = -1.144 + 0.7197x - 0.0379x^2$, $Y = 2.037$, [1.96 ; 2.114]
15.14 (1) - (2) $Y = 68.720 + 0.5664x - 0.007625x^2$, (3) 37.14 , 79.237 [77.147 ; 81.327]
15.15 (1) $Y = 613.56 - 355.11x + 60.4679x^2 - 3.3623x^3$, (2) y = 53,48 , [47.81;59.15]

STIKORDSREGISTER

A
acceptområde 34
additiv model 159, 161
additionssætning
  for sandsynligheder 49
  for linearkomb. af normalf. variable 12, 16
adjusted R^2 196
aftagerens risiko 107
alternativ hypotese 34
approksimation 65, 68
  hypergeometrisk til binomialfordeling 64

B
bagatelgrænse 40, 115
Bartletts test 136, 140
Bayes sætning 51
behandlinger 112
betinget sandsynlighed 50
binomialfordeling
  en variabel 60
  test
    een variabel 64
    en faktor på to niveauer 122
    en faktor på mere end to niveauer 145
  kontrolkort 97
blokforsøg, randomiseret 121, 142, 153
boxplot 91
både A og B 48

C
centrale grænseværdisætning 11
chi i anden fordeling 28
c - kontrolkort 98
Control Chart Sigma 95

D
delforsøg 112
dimensionering 40, 115, 120, 126
diskret variabel 3, 58

E
eksperiment, tilfældigt 2
eksponentialfordeling 77
eksponentiel model 185
ekstrapolation 180
én faktor ad gangen 152
enkelt regressionsanalyse 173
  med gentagelser 189
  uden gentagelser 173
enkelt stikprøvestørrelse 105
    tabel 108
ensfordelte variable 8
ensidet
  binomialtest 64
  chi-i-anden test 39
  Poissontest 68
  t -test 37
  test 34
ensidet variansanalyse 134
  beregninger-formler 135
enten A eller B 48

F
facitliste 216
faktorer 112
  1 faktor på 2 niveauer
    normalfordelte 116
    binomialfordelte 122
    Poissonfordelte 124
  1 faktor på mere end 2 niveauer 133
fakultet 52
fejl af type I 40
fejl af type II 40
fordeling
  binomial- 60
  chi i anden- 28
  eksponential- 77
  hypergeometrisk- 58
  kontinuert 4
  logaritmisk normal- 79
    normal- 10
    rektangulær 76
    t- 23
    Weibull- 79

fordelingsfunktion
kontinuert variabel 6
foreningsmængde 48
forkastelsesområde 34
forklaringsgrad 177 , 180
adjusted 196
formler til beregning af
test af differens mellem
μ_1 og μ_2 i normalfordelte prøver 125
p_1 og p_2 i binomialfordelte prøver 128
μ_1 og μ_2 i Poissonfordelte prøver 128
forudsætninger for
variansanlyse 135
regressionsanalyse 182
F - test 119, 137
fuldstændig randomiseret blokforsøg 142, 163
fællesmængde 48

G
Galton apparat 10
Gauss fordeling 10
generaliseret hypergeometrisk fordeling 80
gennemsnit 14, 20
gentagelser 112
godkendelseskontrol 104
godkendelsestal 105
grænser for kontrolkort 85

H
histogram 4
hovedvirkning 159, 161
hypergeometrisk fordeling 58
hypotesetest
1 normalfordelt variabel 33, 37
binomialfordeling 64
Poissonfordeling 68
middelværdi:2 normalford. stikprøver 116
varians: 2 normalfordelte variable 120
hyppighed, relativ 2
hændelse 2
additionssætning 49
både A og B 48
enten A eller B 48
foreningsmængde 48
fællesmængde 48
ikke A 48
uafhængige 2

I,J
ikke A 48
ikke-parametrisk test 124
inferentiel statistik 1
ISO-9000 83

K
kapabilitet 93
kapabilitetsindeks 94
klyngeudvælgelse 20
kombination 44
kombinatorik 51
konfidensinterval
1 normalfordelt variabel 21, 24, 27
binomialfordeling 64
differens,2 normalfordelte variable 125
differens,2 binomialfordelte variable 128
differens, 2 Poissonfordelte variable 128
for C_p 96
i ensidet variansanalyse 137, 147
i tosidet variansanalyse 155, 159
LSD 138
Poissonfordeling 68
kontinuert stokastisk variabel 3
kontrolkortanalyse 85
kontrol
af binomialfordelt variabel 97
af fejlprocent 95
af normalfordelt variabel 86, 139
af Poissonfordelt variabel 98
løbende 92
varianshomogenitet 136, 146
kontrolkort 85, 91
kontrolgrænser 85
korrelationskoefficient 180
kvadratregel 8
kvalitetskontrol 15
kvalitetsstyring 83
kvalitativ faktor 112
kvantitativ faktor 112

L
leverandørens risiko 107
levetid 78
Levines test 136, 213
linearitetsregel 88
linearkombination 8

lineær model 174
logaritme model 185
logaritmisk normalfordeling 79
løbende kontrol 92

M
middelværdi
  diskret variabel 59, 62
  kontinuert variabel 6
mindste kvadraters metode 175
MR=Moving Range 90
multiplikationsprincip 52
multipel regression 192

N
nedre kontrolkortgrænse 85
nedre kvartil 7
n fakultet 52
niveau for faktor 112
normalfordeling 10
  logaritmisk 79
  normeret 13
  test, 1 variabel 33
    2 variable
      af middelværdi 116
      af varians 119
np - kontrolkort 97
nulhypotese 34

O
opgaver kapitel 3 17
4 31
5 45
6 55
7 71
8 82
9 100
10 110
12 129
13 149
14 169
15 204
oversigt over
  centrale formler i kapitel 4 30
  centrale formler i kapitel 5 43
  centrale formler i kapitel 9 69
  centrale formler i kapitel 11 125
  centrale formler i kapitel 12 147
  fremgangsmåde: tosidet variansanalyse 168
  fremgangsmåde: regressionsanalyse 202
outliers 177

P
parvise observationer 121
Permutation 53
planlægning af forsøg 111, 113, 152
Poissonfordeling 66
  kontrolkort 98
  test
    1 variabel 68
    2 variable 124
    mere end 2 variable 146
polynomialfordeling 80
polynomial regressionsanalyse 196
poolet estimat for varians 136
population 1, 4
potens model 185
produktsætning 49
proces i statistisk kontrol 84
proceskontrol 83
proces ude af statistisk kontrol 84
procesvariablen 86, 97
prædistinationsinterval 25, 183
P-værdi 35
prædistinationsinterval 183

Q
Quality Control 83
Quality management 83

R
randomisering 19, 113
randomiseret forsøg 135
randomiseret blokforsøg 142, 163
rangtest 124
regressionsanalyse
  enkelt
    med gentagelser 189
    uden gentagelser 176

transformation 185
forudsætninger 182
multipel 192
polynomial 196
regressionskoefficienter 174
regressionsligning 174
regressionslinie 174
rektangulær fordeling 76
relativ hyppighed 2
repræsentativ stikprøve 19, 115
R - kontrolkort 86
residual 175
studentized 181
romersk kvadratforsøg 166
rækkefaktor 154

S
SAK = SS 155
sandsynlighed 2, 48
additionssætning 49
betinget 50
produktsætning 49
Satterthwaites test 116
seks sigma proces 94
Shewhart-kontrolkort 85
signifikansniveau 34
simpel udvælgelse 19
s - kontrolkort 86, 90
specifikationsgrænser 93
SPC 83
spredning 6
på gennemsnit 9, 20
SS = Sum of squares 155
statistisk gyldigt forsøg 112
statistisk proceskontrol 83
statistisk uafhængige 2, 49, 112
stikord 211
stikprøve 4 , 19
gennemsnit 14
ordnet 53
spredning 14
størrelse n 85
udvælgelse 19
uordnet 54
varians 6
stikprøvestørrelse 26
stokastisk variabel 3
stratificeret udvælgelse 19
studentized residualer 181
systematisk udvælgelse 19
søjlefaktor 154

T
test
af middelværdi 33, 37
af spredning 39
fejl af type I 40
fejl af type II 40
P-værdi 35
testfunktioner
χ^2 - fordeling 28
t - fordelingen 23
t-fordeling 23
tolerancegrænser 93
tosidet test 38
tosidet variansanlyse 152, 156
transformation 185
Trinvis variansanalyse 164
t-test, ensidet 37
tæthedsfunktion
diskret variabel 59, 62
kontinuert variabel 4, 6

U
uafhængige hændelser 2 , 49
uafhængige statistiske variable 2,8, 112
ubalancerede forsøg 164
udfald 2
udfaldsrum 2
uordnet stikprøveudtagning 54

V
variabel
binomialfordelt 60
diskret 3, 58
kontinuert 3
stokastisk 3
varians
diskret variabel 6
kontinuert variabel 6
variansanlyse
ensidet 134
beregninger, formler 135
tosidet 152

beregninger, formler 154
med vekselvirkning 156
additiv 154, 159, 161
varianshomogenitet 136 ,182, 190
vekselvirkning 154, 156

W
Weibullfordeling 79

X
x - streg kontrolkort 86

Z
Z - fordeling 13

Ø
øvre kvartil 7
øvre kontrolgrænse 85