Statistik i grænseregionen

Kompendium til statistik I og II, BA int. Flensborg

Sørensen, Nils Karl

Publication date:
2013

Document version:
Indsendt manuskript

Citation for pulished version (APA):
Sørensen, N. K. (2013). *Statistik i grænseregionen: Kompendium til statistik I og II, BA int. Flensborg*. Institut for Grænseregionsforskning – Syddansk Universitet.

Nils Karl Sørensen ©
Statistik I og II, BA int, Flensborg
Udgave 2012/2013

Statistik i grænseregionen

Kompendium til statistik I og II

BA int. Flensborg

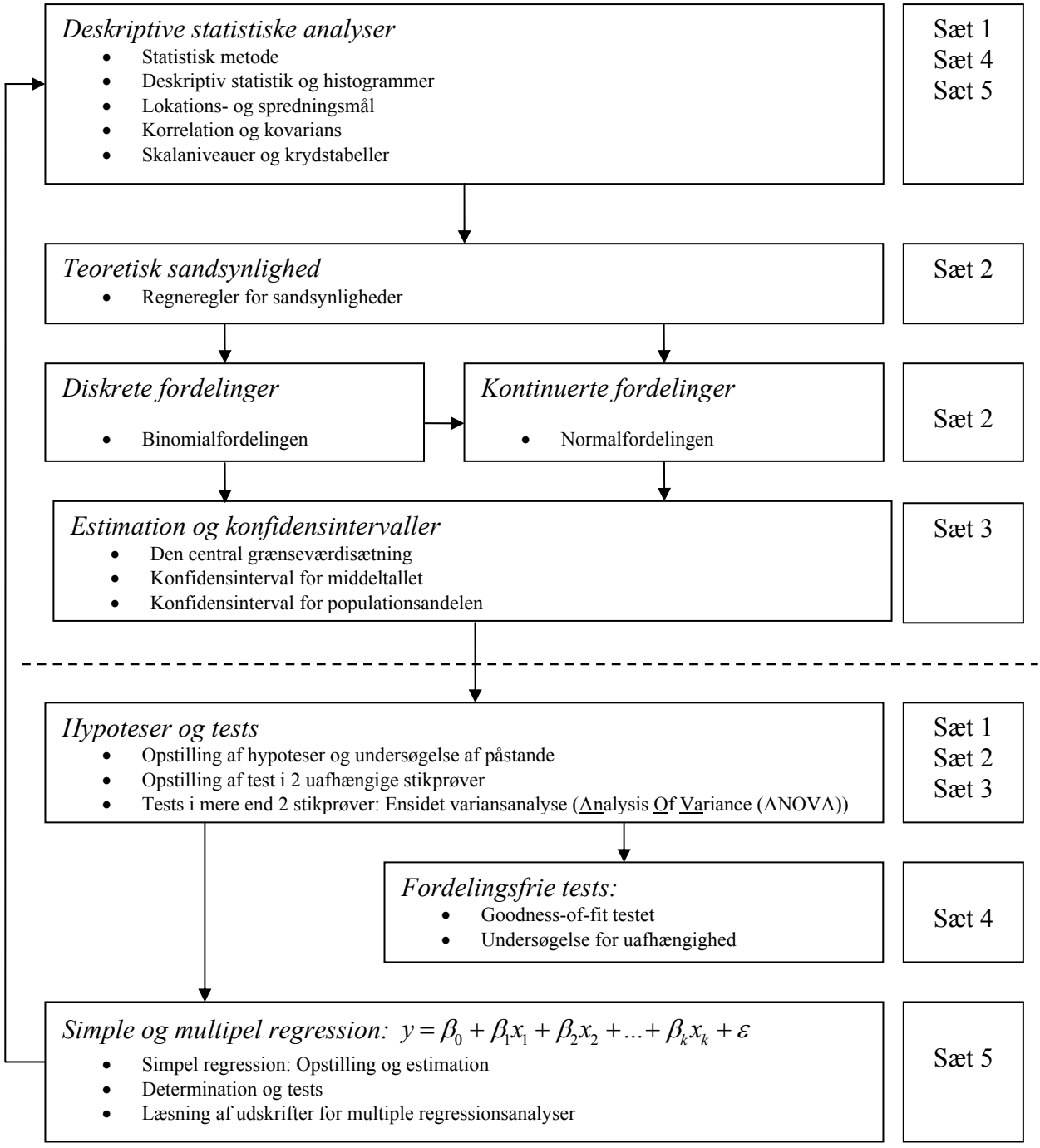
1. Indledning og emnekredse

Disse noter omfatter den teoretiske referenceramme til undervisningen i statistik I og II ved BA int studiet i Flensborg. Kurset omfatter to undervisningsforløb, hver af cirka 12 uger med to lektioner per uge. Foruden noterne er der øvelsesopgaver. Disse løses, som en integreret del af undervisningen. Noterne er tilrettelagt således, at disse modsvarer den undervisning, der gives på det tysksprogede hold. Derfor titlen på det samlede kompendium "statistik i grænseregionen"! God læse- og arbejdslyst ☺

Dertil kommer statistiske tabeller samt øvelsesopgaver

Skematisk opstilling af kurserne statistik I og II

Den øverste del over den stiplede linje giver emnekredsene i Statistik I, mens den nederste del giver emnekredsene i Statistik II. Det ses, at emnekredsene ikke er uafhængige af hinanden. Meget groft sagt, så er den deskriptive statistik en analyse af udseende af fordelingen af et enkelte uafhængige datasæt. Emnekredsene tilstræber, at der ved anvendelse af statistiske teorier sker en integration således at de enkelte dataserier kan relateres til hinanden.



Nils Karl Sørensen ©
Statistik I, BA int Flensborg

Sæt 1: Statistisk metode og deskriptiv statistik

af Nils Karl Sørensen

Indhold *side*

Bemærkning
I alle noter til såvel Statistik I som til Statistik II anvendes den engelsk/amerikanske notation for tusind- og kommaseparator. Det vil sige, at eksempelvis éttusindefirehundrede skrives 1,400 og to komma tre skrives som 2.3. Dette er indført for at lette notationen til lommeregnere, hvor denne form for notation er den mest brugte.

1. Statistisk metode

Jeg har hovedet fyldt med statistik, for jeg har bemærket, at man ikke kan bevise noget uden statistik

Mark Twain

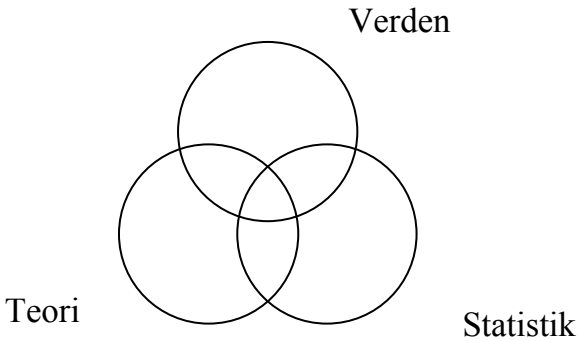
I alle medier bombarderes vi med informationer i form af tal ofte i form af grafikker, gennemsnit og meget andet! Selv vejrudsigten kræver, at man kan læse statistisk materiale.

Ordet *statistik* kommer fra det latinske *statisticum collegium* (statsrådgiver) og det italienske ord *statista* (statsmand eller politiker). Ordet kom til Danmark via tysk indføring af Gottfried Achenwall i 1749, og beskrev oprindeligt behandling af data for staten. Dengang var det mest statistik om eksempelvis befolkningens sammensætning militære hemmeligheder. *Statistik* er i dag en videnskabelig metode, der baserer sig på anvendelse af numeriske tal. Ofte vil man bearbejde materiale fra statistikbanker eller spørgeskemaundersøgelser for at bekræftet eller forkastet hypoteser, der tager sit udspring for eksempel i økonomisk teori.

Analytiske arbejdsprocesser tager sit udspring i følgende forhold:

- Virkeligheden (som den observeres)
- Teorien (baseret på observationer formuleres teorier)
- Statistikken (baseret på registreringer af observationer)

Sammenhængen mellem de tre forhold kan anskueliggøres i diagrammet:



Gode analyser indeholder elementer fra alle forholdene, og vil derfor finde sted i den lille fællesmængde mellem de tre mængder. Bemærk at der er mange fejlkilder! Observationer af fænomener skaber forundring, som er en inspiration til opstilling af teorier. Statistik er en

vej til at vise, om teorierne er valide eller ej. Statistik eller en anden metode er således et bindeled eller en integreret del af en analytisk arbejdsproces.
Astronomen Tycho Brahe (1546−1601) var en af de første, der indså denne sammenhæng. I 1572 observerede han en ny og meget stærk stjerne på himlen (det, som nu kaldes en Supernova). På den tid havde det været læren, at himlen var en konstant og uforanderlig størrelse. Denne observation ledte Tycho Brahe til den overvejelse, at himlen ændres over tid. For at verificere denne teori, var det nødvendigt at foretage systematiske observationer nat efter nat af bevægelser på himlen. Dette ledte til statistikker over bevægelserne på himlen, som kunne anvendes til videre analyser og opstilling af hypoteser. Det vil sige en direkte anskueliggørelse af den postulerede sammenhæng i figuren ovenfor.

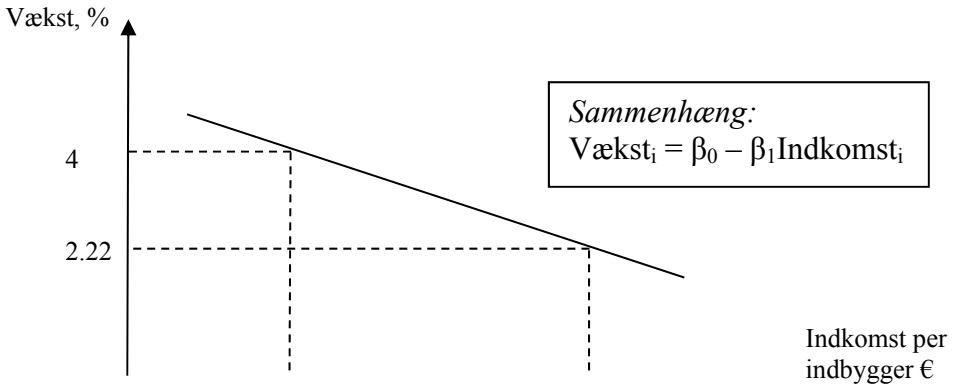
For at man kan anvende *statistisk materiale* på denne måde, må man kunne gøre følgende:

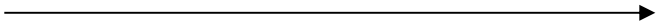
1. Teorien må kunne *operationaliseres.* Det vil sige, at den kan omsættes til målbare data
2. Det må være muligt at finde eller indsamle statistisk materiale, som reflekterer teorien

Det er vigtigt at gøre sig klart, at der i samspillet mellem teori og statistik, tillige er et samspil mellem to forskellige sæt af hypoteser, der særskilt skal undersøges. Nemlig de teoretiske og de statistiske. Et eksempel kan måske anskueliggøre dette.

Diskussion af en analyse af den regionale vækst i Tyskland
Antag man ønsker at undersøge regionale forskelle i økonomisk vækst i Tyskland. Som en følge af forskelle i sammensætningen af kapital og arbejdskraft fra den ene region til den anden, vil der være forskelle i vækst og udvikling. Eksempelvis er indkomsten per indbygger højere i Hamburg end i Flensburg. Det betyder også, at væksten i Flensburg bliver den højeste. Antag eksempelvis, at alle indkomster i hele Tyskland øges med 1,000 €. Er indkomsten per indbygger i Hamburg på 45,000 €, vil stigningen være på 2.22 procent, mens stigningen i Flensburg, hvor indkomsten per indbygger er på 25,000 €, vil være på 4 procent.

På denne baggrund kan man forestille sig følgende sammenhæng mellem den økonomiske vækst og niveauet for indkomsten:





Her er de to observationer indtegnet og forbundet med en linje. På bagrund af informationerne, kan der formuleres en sammenhæng, der siger, at den økonomiske vækst i de mindre velstående regioner som regel vil være højere end i de mere velstående regioner. Når væksten i regioner med lav indkomst, er den højeste, så vil indkomsterne i Flensburg og Hamburg over tiden tilnærme sig hinanden. Dette kaldes *den økonomiske konvergensteori.*

Med udgangspunkt i observationerne kan der anvendes analytiske metoder fra eksempelvis mikroøkonomi (VWL I) og makroøkonomi (VWL II) til at formulere en sådan model for konvergens.

Modellens validitet skal nu undersøges. Til det formål skal der anvendes et statistisk materiale. Statistik om indkomsterne i regionerne i Tyskland for en given periode kan findes enten ved Statistisches Bundesamt eller ved EUROSTAT.

Nu skal den statistiske teori finde anvendelse. I figuren er gengivet en formel for en linje af formen:

$$Vækst_i = \beta_0 - \beta_1 Indkomst_i \qquad i = 1,2, \ldots, n$$

Linjen omfatter alle *n* regioner i Tyskland, mens *i* er tælleenheden. Der er et konstantled β_0 og en hældning β_1. Modellen siger, at hældningen giver *konvergensgraden.* Den matematiske linje kan udledes fra en model, som bruger mikro- og makroøkonomi. Med data kan man så beregne værdier for konstantled og hældning.

Den statistiske metode til at beregne ligningens værdier kaldes *lineær regression* og vil blive behandlet i statistik II. For at denne metode skal kunne give det korrekte resultat, skal den række forudsætninger til den statistiske model være opfyldt. Materialet skal eksempelvis følge Normalfordelingen, der er den hyppigst anvendte statistiske fordeling. Hvis denne forudsætning ikke er opfyldt vil beregningen af hældningskoefficienten blive misvisende, og informationerne om konvergensgraden vil blive ubrugelige.

Eksemplet søger at illustrere, at arbejdet med statistik er en del af en integreret proces, hvor der er et samspil mellem virkelighed, den økonomiske teori og den statiske metode. Hvis der er noget, som går galt i processen, bliver resultatet misvisende. Det er lidt, som at skulle lande en rumraket på jorden. Her er indfaldsvinklen afgørende. Hvis den er for stor, så brænder rumskibet op ved mødet med atmosfæren; hvis vinklen er for lille, så smutter rumskibet ud i rummet igen og flyver væk fra Jorden. Kun hvis vinklen er korrekt, kommer rumskibet sikkert ned på jorden!

En integreret analysemetode

Deskriptiv statistik, der er emnet for denne note, udgør fundamentet for *statistisk inferens*, der er emnet for de øvrige noter og hele kurset Statistik II. Sidstnævnte indeholder emner som sandsynligheder og –fordelinger, undersøgelser af hypoteser med mere. *Statistisk*

inferens kan oversættes som *logisk slutning.* Det vil sige kunsten at udlede konklusioner om en totalpopulation ud fra en stikprøve.
Der kan opstilles en analysemetode til udarbejdelser af eksempelvis rapporter, som er baseret på disse elementer. Denne består af:

- Indsamling af statistisk materiale eksempelvis fra databanker eller spørgeskemaer
- Præsentation af det statistiske materiale i form af tabeller og figurer
- Forberedelse af analysen:
    - Opstilling og formulering af hypoteser
    - Valg og opstilling af en statistisk model
    - Estimation af den statistiske model
    - Evaluering af modellen og dens underliggende statistiske forudsætninger
- Relevante konklusioner der kan udledes fra materialet i forhold til den opstillede problemformulering

2. Brug af lommeregner og computer i statistik

Kurser i statistik har undergået væsentlige forandringer igennem de seneste 30 år. Tidligere blev der anvendt mange ressourcer på matematisk programmering, så man kunne beregne de værdier af tests med mere, som er en del af den statistiske metode.

Fremkomsten og billiggørelsen af lommeregnere samt menu styret programmel har ændret dette forhold, og har lettet regnearbejdet. Undervisningen i statistik er blevet mere analytisk præget end tidligere, og dette er en god ting! Nu får man i højere grad præsenteret materiale i form af udskrifter, hvor man skal forholde sig dels til beregningsmetoden dels til en fortolkning af udskriften eller et skærmbillede. Det betyder også, at eksamen i højere grad er problemorienteret. Ud fra en nogle oplysninger skal der vælges en metode, hvorefter der opstilles og udregnes værdier. Endelig skal resultatet kommenteres.

Hele dette kursus, samt kurset statistik II kan laves på en **lommeregner**. Imidlertid vil det også blive vist, hvordan man foretager mange af tingene i regnearket **Excel**, der er en del af Microsoft Office pakken. **Excel er ikke et eksamenskrav**, men vises, da man nok på et eller andet tidspunkt i sin studietid får behov for at arbejde med data for eksempel i forbindelse med udarbejdelse af større opgaver

Valg og brug af lommeregner

Hvis man ikke har en lommeregner, så er det nødvendigt at anskaffe sig en sådan (ordet ”lommeregner” er misvisende – se billederne på næste side)! Foruden til statistikkurserne, så kan denne også anvendes til eksempelvis finansiering og mikroøkonomi.

På næste side findes et billede af de to lommeregnere, som forfatteren til disse noter har anskaffet. Til venstre ses en **Texas TI-84 Plus** og til højre en **Texas IT-89 Titanium.** Sidstnævnte findes også i en variant med tastatur kaldet **Voyage 200.** Alle lommeregnere

kan bruges til dette kursus, men **jeg vil på det kraftigste anbefale, at man anskaffer en Texas TI-84 Plus.** Årsagerne hertil er følgende:

- Den koster mindre, og kan alt, hvad man skal bruge
- Skærmen er lettere at læse på TI-84eren. Det står MEGET småt på TI-89eren
- Menu systemet er lettere at anvende på TI-84eren

I disse noter vil det generelt blive vist, hvordan man løser problemstillingerne på Texas **TI-84 lommeregneren**. På de to andre lommeregnere er det næsten det samme, men måden man starter op på, er lidt anderledes. Det forklares i det følgende.

Det er vigtigt at man træner sig i brugen af lommeregneren, da der ved anvendelse af denne kan spares meget tid! Der skal også henvises til **manualen til lommeregneren!** Det er nyttig læsning, og selv om underviseren er en vidende person, så ved denne ikke alt (vi er jo trods alt ikke guder)!

Har man en lommeregner af andet fabrikat eksempelvis HP eller en ældre type af en Texas, så henvises til manualen. Meget af teknologien i såvel lommeregnere fra HP og Texas blev udviklet i 1980erne, og er ikke ændret væsentligt siden. Undertegnes TI-30X fra 1992 kan eksempelvis det samme som mine børns lommeregnere, der hedder TI-30BX Multiview, men især brugervenlighed og display er forbedret i den nye udgave.

Se man først på **TI-84eren**, så skal man især bruge tasterne STAT og DISTR. Sidstnævnte fås ved anvendelse ad 2ND → DISTR.

På **TI-89eren** er det lidt anderledes, som det måske kan anes på illustrationen ovenfor til højre. Der vælges i den overordnede menu STATS/LIST E → ENTER (kommer der herefter noget om menuer, så vælges OK=ENTER). I det skærmbillede, der vises, kan man vælge funktioner med F-testerne. Eksempelvis fås DISTR ved at vælge F5. Med piletasterne øverst til højre kan man gå rundt i F-menuerne. I skærmbilledet vil det i toppen blive vist, hvilke funktioner der er tilknyttet de fem F-taster.

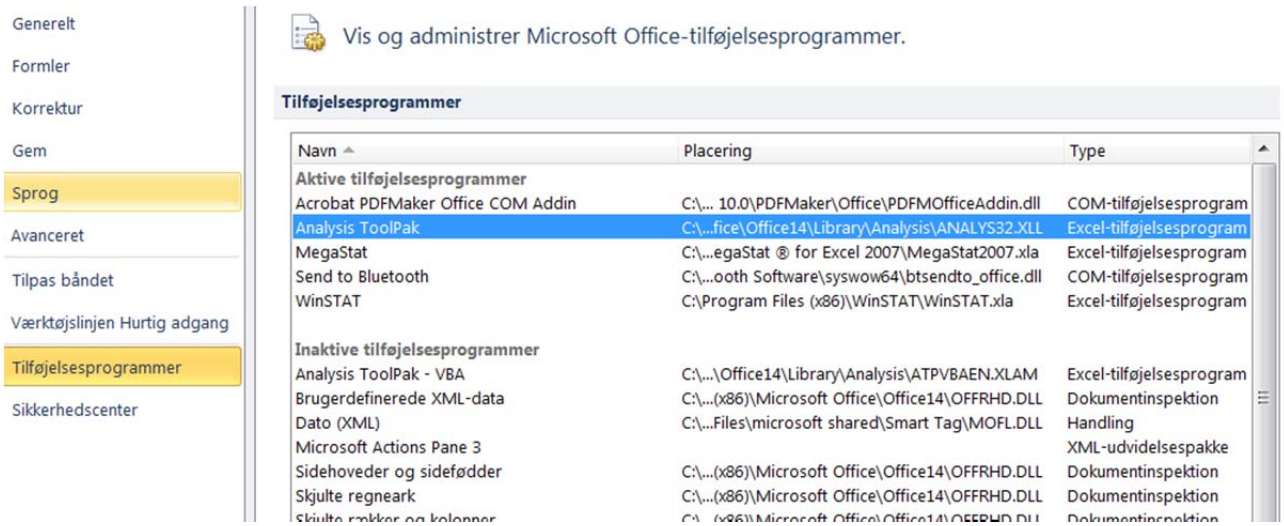
I undermenuerne er der ingen forskel på de to lommeregnere!

Brug af computer

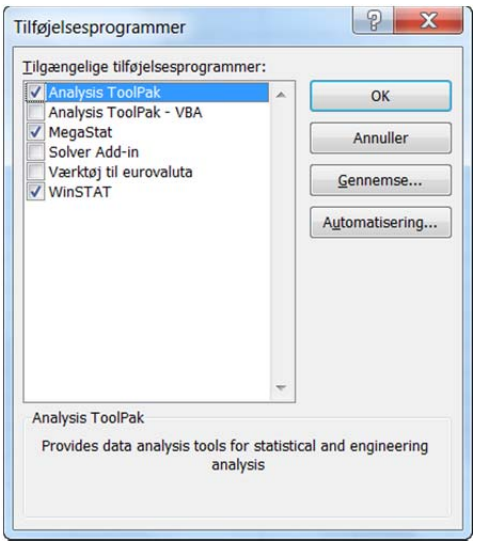
Microsoft Office pakken findes både til Windows og til MAC, og skal være installeret. Metoden til installation i Excel 2007, 2010 og 2011 er den samme.

For at kunne arbejde med statistisk analyse skal **Analysis ToolPak** være aktiv. Denne indeholder en række analytiske værktøjer til statistisk analyse. Pakken gøres aktiv på følgende måde:

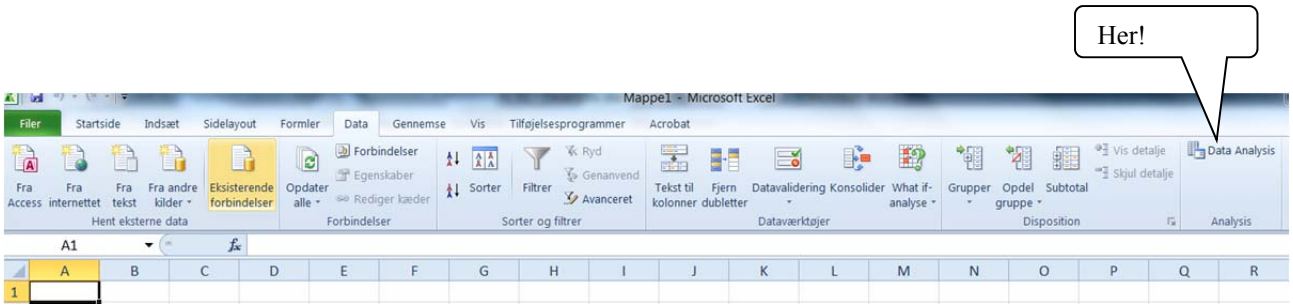
Åben **Excel,** og gå ind inder fanen FILER. Her vælges TILFØJELSESPROGRAMMER og så INDSTILLINGER.



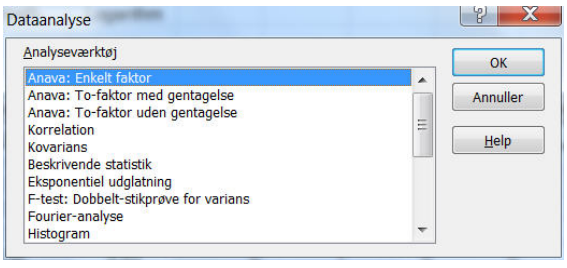
Her markeres **Analysis ToolPak** og i bunden af menuen trykkes der på UDFØR. Nu fremkommer følgende billede:



Der sættes en markering ved **Analysis ToolPak** og trykkes OK. Nu skulle der under fanen DATA i Excel være et element længst til højre med titlen **Data Analysis** jævnfør illustrationen nedenfor.



Når man trykker på **Data Analysis**, så kommer følgende menu:



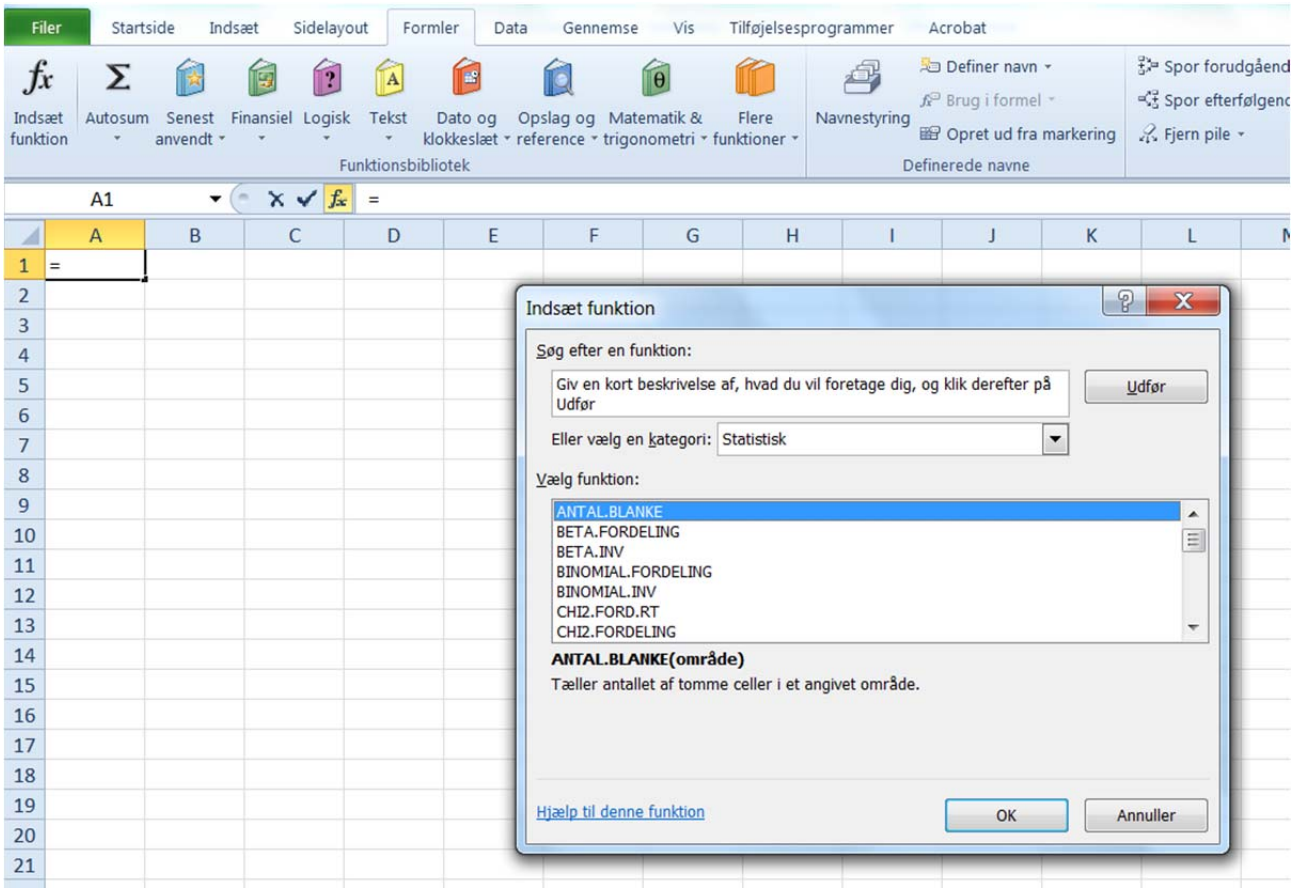
(Nogle steder vil menuen være på engelsk, selvom man har valgt dansk. Undertegnede er svar skyldig om årsagen).

Til undermenuerne findes der en udmærket hjælpefunktion med forklaringer af de analytiske redskaber.

Andre muligheder i Excel

Foruden **Analysis Toolpak** findes der i **Excel** også andre muligheder for at arbejde med statistik. Under fanen FORMLER findes menuen **Indsæt funktion.** Trykker man på denne fremkommer den menu, der er gengivet i skærmbilledet. Under "vælg en kategori" kan der vælges den ønskede funktion.

Denne funktion vil i kurset blive anvendt til at finde eksempelvis midtpunktet for en fordeling og sandsynligheder udledt fra de teoretiske fordelinger. Endelig er funktionen anvendt til at opstille de tabeller, som findes i tabelværket **Statistics Tables**, som senere vil blive anvendt i kurset. **Statistics Tables** findes til download i Blackboard.



Specielt for MAC-brugere

På Office pakken 2010 og 2011 til Mac er Analysis ToolPak blevet droppet af uransagelige årsager, og jeg er forhindret i at spørge Steve Jobs om årsagen! Imidlertid skal man ikke fortvivle, da der er lavet servicepakke med det nødvendige. Denne finder man til download på adressen: http://www.microsoft.com/en-us/download/details.aspx?id=17198

Installationen er da som ovenfor! For Office 2012 og 2013 kender jeg ikke løsningen ☹

3. Hvad er deskriptiv statistik?

I *deskriptiv statistik* undersøges udseendet af et datasæt, og der fremdrages en række karakteristika ved datasættet. Det kan være den typiske værdi, datasættes spredning, skævhed med mere.

Formålet med at udarbejde en deskriptiv statistisk undersøgelse er, at opnå kendskab til den underliggende data genererende proces, også kaldet DGP. På denne baggrund kan man eksempelvis fastlægge, om datasættet følger en given teoretisk statistisk fordeling. Dette er en god viden at have, hvis der skal opstilles nogle generelle retningslinjer for at arbejde med datasættet.

En deskriptiv statistisk undersøgelse af et datasæt består af tre elementer:

- En grafisk præsentation af materialet eksempelvis i form af et *histogram*
- En præsentation af materialets positionsmål = hvad er det typiske?
- En præsentation af materialets spredningsmål = hvor stor er usikkerheden?

Statistisk materiale kan klassificeres for forskellig måde. Normalt skelnes der mellem følgende *typer* af data:

- Tværsnit: Et antal sektorer/kategorier/regioner på et givet tidspunkt
- Tidsserie: En variabel over en tidsperiode: år, kvartaler eller måneder
- Panel: En kombination af tværsnits- og tidsseriedata
- Census: Statistisk materiale fremkommet fra et spørgeskema

Der kan forekomme overlap mellem de fire typer. I eksemplet om den lille regionale konvergensmodel for Tyskland i det første afsnit af disse noter, er der tale om tværsnitsdata. Men da den økonomiske vækst anskues over tid, må materiale også være et tidsserie datasæt.

Paneldata er specielle. Her følges typisk et forløb over tid. Eksempelvis kan en person have skiftet arbejde fem gange på ti år. Her bruger man data til eksempelvis at beregne sandsynligheden for, at personen vil skifte arbejde igen opgjort i forhold til en række baggrundsvariable som for eksempel indkomst og uddannelse. Sådanne data anvendes ofte i sociologiske sammenhænge og undersøgelser af forhold på arbejdsmarkedet.

4. Grafiske præsentationer og histogrammer

En grafisk præsentation anvendes til at give et bedre overblik af fordelingen af et materiale. Ofte vil det være hensigtsmæssigt at sammentælle data. Dette gøres ved at beregne *frekvensen*. Frekvensen er lig med hyppigheden. Det vil sige, hvor ofte en observation forekommer.

Et *histogram* er en grafisk illustration af en frekvensfordeling. Normalt anvendes den lodrette akse (y-aksen) til at vise frekvensen, mens den vandrette akse (x-aksen) anvendes til at angive data. Dette kaldes ofte for referencepunkt eller intervalskalaen afhængigt af materialet. I notesæt 5 vendes der tilbage til emnet om skalaer.

Eksempel på histogram med intervalskala

Betragt et statistisk materiale med 20 observationer af månedlige indkomster opgjort i 1,000 DKK. I tabelform ser det ud som følger:

9	6	12	10	13	15	16	14	14	16	17	16	24	21	22	18	19	18	20	17

Det vil sige at n=20. Det ser lidt uorganiseret ud, så for at lette overskueligheden ordnes data, så den mindste observation kommer først.

6	9	10	12	13	14	14	15	16	16	16	17	17	18	18	19	20	21	22	24

Det næste interessante forhold er, at der er nogle af observationerne, der optræder mere end en enkelt gang. For eksempel observeres ”16” hele 3 gange. Denne observation har således hyppigheden 3. Det er præcis denne viden, der anvendes, når der opstilles et *histogram.*

Hvordan kan materialet fordeles på nogen hensigtsmæssige grupper? I den følgende tabel er materialet søgt opdelt i grupper med en intervalbredde på 5 ved anvendelse af ”pile” eller markeringer for hver observation.

	Under 5	6 til 10	11 til 15	16 til 20	21 eller mere	Total
Antal		|||	|||||	|||||||||	|||	20
Frekvens	0	3	5	9	3	20
Relativ %	0	0.15	0.25	0.45	0.15	1.00
Kumulativ %	0	0.15	0.40	0.85	1.00	

I de to nederste linjer findes et par yderligere beregninger. I den tredje linje findes den *relative frekvens*. Den er defineret som $f_i = x_i/n$, hvor x er antallet af observationer, som findes indenfor intervallet i. Endelig findes i den fjerde linje den *kumulative frekvens*. Det er den summerede værdi af de relative frekvenser. Den kumulative frekvens summerer til 1.00.

Hvordan finder man den rigtige størrelse eller bredde af et interval? Som oftest prøver man sig frem, men der findes også en mere matematisk baseret metode. Her anvendes formlen $2^k=n$, hvor k er lig med antallet af intervaller. Da $2^4=16$ vælges i eksemplet $k=4$.

Bredden af intervallet kan nu findes ved anvendelse af følgende formel:

$$\frac{(x_{\max} - x_{\min})}{k} = \frac{24-6}{4} = 4.5$$

Intervallerne bliver nu som [6 til 10.5[; [10.5 til 15[; [15 til 19.5[og [19.5 to 24]

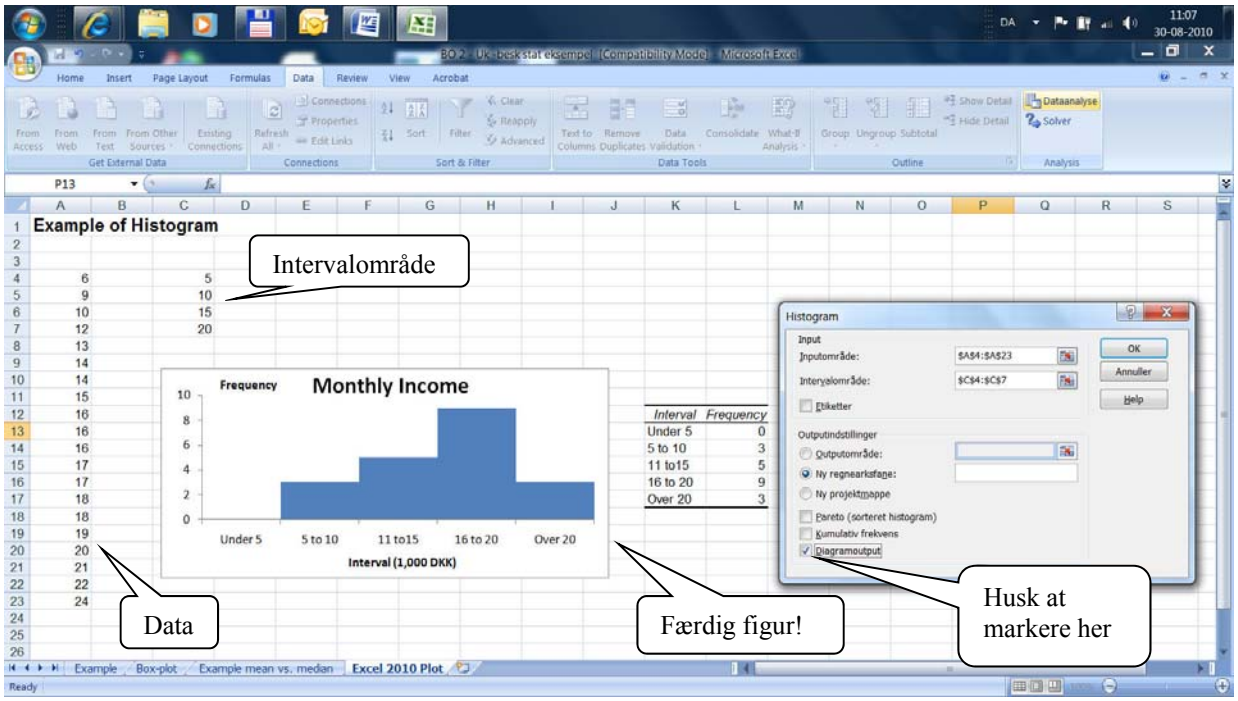
Tabellen ser nu ud som:

	10.5 til 15	16 til 15	15 til 19.5	19.5 til 24	Total
Antal	||| ||||| ||||||| ||||	20
Frekvens	3	5	8	4	20
Relativ %	0.15	0.25	0.40	0.20	1.00
Kumulativ %	0.15	0.40	0.80	1.00	

Histogrammerne vil blive næsten identiske, uafhængigt af hvilken fremgangsmåde, der er valgt!

Tegning af histogram i Excel

Det letteste i eksamenssituationen er at lave en tegning i hånden af histogrammet. Er der bedre tid til rådighed, udarbejdes histogrammet med anvendelse af **Excel.** Dette gøres som følger ved i DATA at vælge DATA ANALYSIS, og så vælge HISTOGRAM. Dette er gjort på skærmbilledet nedenfor.



I dialogboksen HISTOGRAM skal man vælge imputområdet med data. Dernæst vælges et intervalområde. Intervalområdet skal lave separat. Dette er også vist ovenfor. Intervallerne er den nedre grænse. Husk at markere om der er eventuelle etiketter, og at udskriften ønskes om et diagram. I den endelig figur er der blevet rettet lidt til for at få en pæn figur. Især er figurens plads søgt at gøre så stor som muligt.

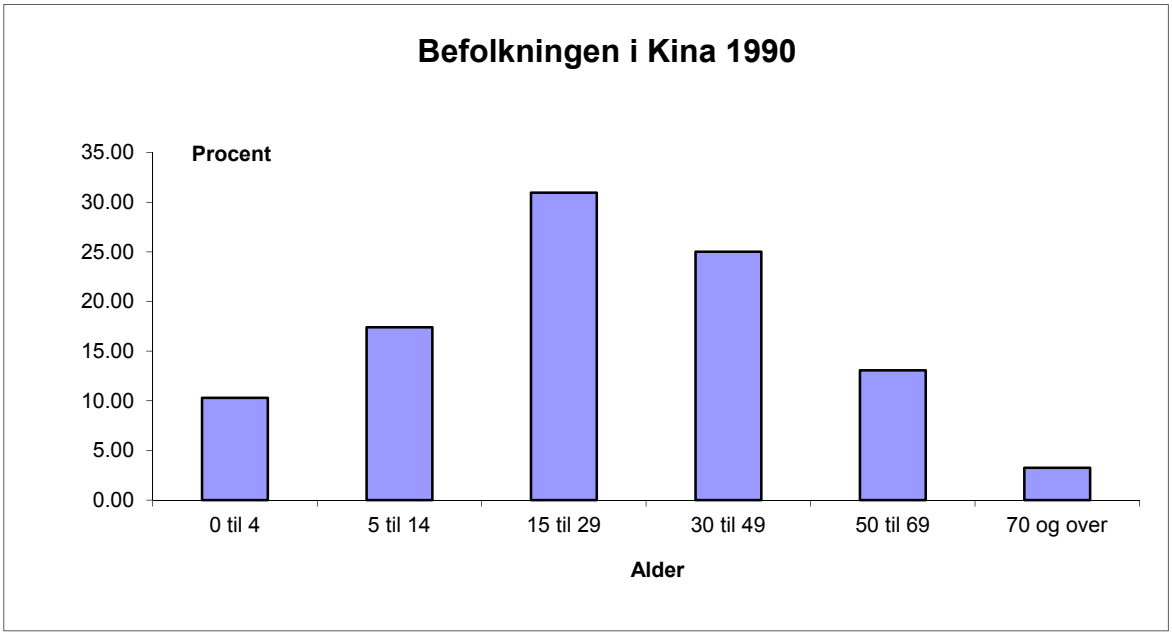
Histogram i Excel med varierende intervalbredde

Tabellen viser fordelingen af befolkningen i Kina, som den så ud per første juli 1990. Foruden absolutte data vises også den relative fordeling.

Alder, år:	**0 til 4**	**5 til 14**	**15 til 29**	**30 til 49**	**50 til 69**	**70 og over**	**I alt**
Personer, millioner	116.60	196.90	350.50	283.10	147.90	36.80	1131.90
Personer, %	10.30	17.40	30.97	25.01	13.07	3.25	100.00
Enheder á 5 år	1	2	3	4	4	[4]*	18
% enheder á 5 år	10.30	8.70	10.32	6.25	3.27	0.81	

*=antaget, se teksten

Tegnes et søjlediagram i Excel med ovennævnte data, som i anden linje kan det se ud som følger:



Det fremgår af diagrammet, at de fleste kinesere i 1990 var at finde i aldersintervallet mellem 15 til 29 år. **Illustrationen er imidlertid misvisende!** Dette skyldes den varierende intervalbredde. Eksempelvis er bredden i det første interval lig med 5 år, mens den i intervallet mellem 5 og 29 år er lig med 15 år!

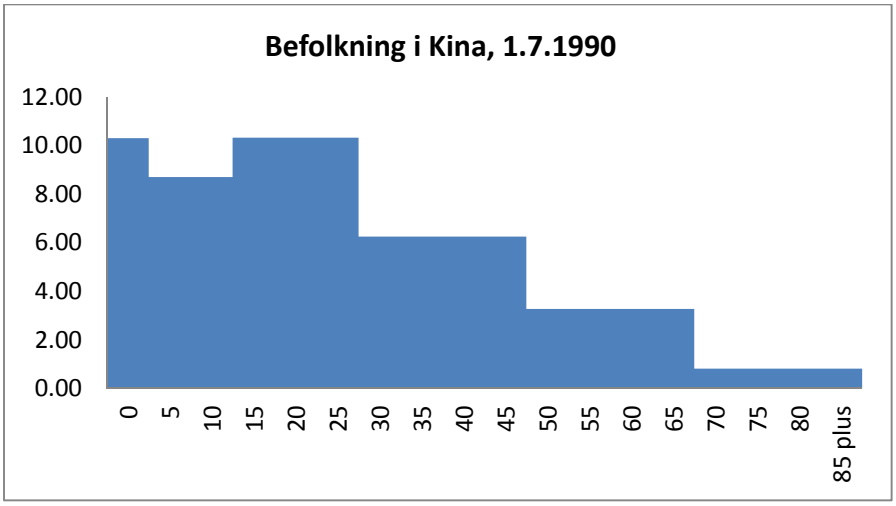
Problemet med den varierende intervalbredde skal selvfølgelig indarbejdes i præsentationen. Dette gøres ved at opdele i mindre enheder med identisk bredde af intervaller. I den tredje linje i tabellen er gengivet antallet af "enheder à 5 år". Denne målestok er valgt, da den er lig med det mindste interval, der er anvendt i statistikken. For den første aldersgruppe er der én enhed, mens den i den næste er to enheder og så fremdeles. For intervallet over 70 år er

det lidt specielt. Her har forfatteren antaget en bredde på intervallet på 4 enheder svarende til 20 år. Forfatteren antager således, at der i 1990 stort set ikke fandtes kinesere på 90 år og derover. Denne antagelse er naturligvis subjektiv og åben for diskussion. Vurderingen er foretaget med baggrund i den turbulente historie i Kina i det forrige århundrede! I den fjerde linje er den procentuelle fordeling fra anden linje divideret med antallet af "enheder á 5 år" fra tredje linje.

Nu skal figuren tegnes. Dette gøres ved at konstruere nedenstående tabel, hvor data er "skrevet ud":

Alder, år	0	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	>85
Person,%	10.3	8.7	8.7	10.3	10.3	10.3	6.25	6.25	6.25	6.25	3.3	3.3	3.3	3.3	0.8	0.8	0.8	0.8	0.8

Nu kan figuren (eller histogrammet) let tegnes som et simpelt søjlediagram (i 2-D). Med lidt efterbehandling ser det korrekt histogram ud som følger:



Sammenlignes de to præsentationer, så fremgår det, at konklusionen er ændret. Nu er det både intervallet "0 til 5 år" og "intervallet "15 til 29 år", som er de typiske intervaller. Den første præsentation overdriver således effekten af "et-barns politikken" som blev indført i 1970erne.

Problemstillingen med den varierende intervalbredde forekommer oftere end man tror det og oftest i arbejdsmarkeds- og befolkningsstatistikker. I "Statistisk Tiårsoversigt" som udgives af Danmarks Statistik findes der en tabel over "landbrugsbedrifter med tilhørende areal opgjort efter størrelse" med tilhørende figur. I den seneste udgave er denne figur misvisende. Dette er imidlertid ikke en fejl af ny dato. Da forfatteren til disse noter selv var student i slutningen af 1970erne, var det den sammen fejlagtige illustration, der blev vist i "Statistisk Tiårsoversigt"!

5. Lokationsmål

Lokationsmål kaldes også positionsmål og betegner udtryk for det oftest observerede. Årsagen, til at der anvendes forskellige lokationsmål, er, at de forskellige mål inddrager forskellige mængder af information. Nogle af lokationsmålene skulle gerne være kendt fra folkeskolen eller gymnasiet.

I det følgende omtales:

- Middeltallet
- Modus eller typetallet
- Medianen
- Kvartiler og percentiler
- Det geometriske middeltal

Middeltallet

Ved beregning af middeltallet inkluderes alle observationer i et datasæt. Middeltallet er det, som man i daglig tale kalder "gennemsnittet". For en totalpopulation betegnes middeltallet som μ, mens middeltallet i en stikprøve betegnes $\overline{X}$. Følgende formler kan anvendes til at beregne middeltallet:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

For en *grupperet fordeling* med k grupper med frekvensen f_i haves:

$$\overline{X} = \frac{\sum_{i=1}^{k} f_i \times x_i}{n}$$

Betragt følgende eksempel. Middeltallet for det datasæt om indkomster, som der blev tegnet et histogram for i det forrige afsnit, findes som:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{6 + 9 + 10 + ... + 21 + 22 + 24}{20} = \frac{317}{20} = 15.85$$

Betragt, som et eksempel på en grupperet fordeling, et datasæt, der viser resultaterne af bestående studenter, fra eksamen i VWL-III også kaldet International Økonomi fra BA int. i Flensburg fra februar 2011. Fordelingen var givet som:

Karakter	2	4	7	10	12	Total
Frekvens	10	26	33	19	4	92

(Det skal bemærkes at der var 8 studenter, der dumpede, så andelen af bestående var fin) ☺

Gennemsnittet beregnes nu som:

$$\overline{X} = \frac{\sum_{i=1}^{k} f_i \times x_i}{n} = \frac{10 \times 2 + 26 \times 4 + 33 \times 7 + 19 \times 10 + 4 \times 12}{92} = \frac{593}{92} = 6.45$$

Modus eller typetallet

Modus eller *typetallet* er den observation, der optræder hyppigst. Det vil sige, den observation med den største frekvens.

I eksemplet med indkomsterne er modus lig med 16, der observeres 3 gange, mens modus i eksemplet med fordelingen af karakterer lig med 7, som observeres 33 gange.

Medianen

Medianen er den midterste observation i datasættet, når observationerne er opstillet efter størrelse. Medianen findes matematisk ved anvendelse af udtrykket:

Medianen = 0.50(n + 1) ordnede position

I eksemplet med indkomsterne er der 20 observationer. Medianen findes således ved observation nummer 0.50(20+1) = 10.5. Tælles der mod højre i tabellen nedenfor ses det, at medianen er lig med 16. Såvel observation nummer 10 som nummer 11 antager denne værdi. Medianen er illustreret med farvet signatur.

I eksemplet med fordelingen af karakterer er der 92 observationer, så medianen findes ved observation nummer 46.5. Denne er lig med værdien 7.

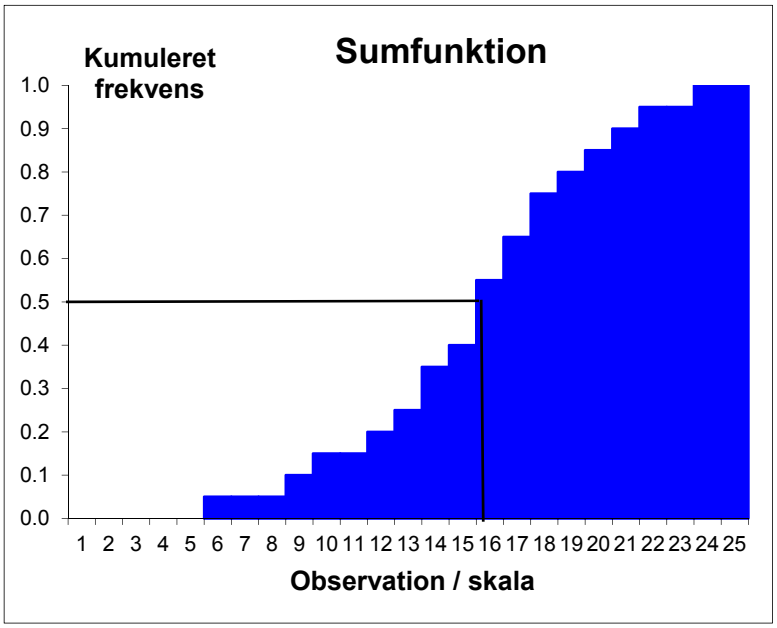
Data	6	9	10	12	13	14	14	15	16	16	16	17	17	18	18	19	20	21	22	24
Frekvens	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05
Kumuleret	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95	1.00
Nummer	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*

Fælles for modus og medianen er, at ikke alle data, der er tilgængelige. Ved beregningen af middeltallet indgår alle observationer. For medianen betyder det, at ekstremer ikke øver indflydelse på dennes værdi. Dette gør medianen til en velegnet parameter i lønstatistikker, hvor det netop er den typiske løn, der forhandles om, og ikke enkelte meget høje lønninger. Der vendes tilbage til dette i det sidste afsnit i disse noter.

I tabellen er der tillige to beregninger dels frekvensen for den enkelte observation defineret som $1/n$, dels den kumulerede frekvens. Sammenstilles den kumulerede frekvens og observationerne fås ***sumfunktionen.*** Denne er vist i figuren, hvor medianen også er indtegnet.

Sumfunktion for 20 indkomster i 1,000 DKK

Obs	Frekvens	Kumuleret frekvens
1	0	0.00
2	0	0.00
3	0	0.00
4	0	0.00
5	0	0.00
6	1	0.05
7	0	0.05
8	0	0.05
9	1	0.10
10	1	0.15
11	0	0.15
12	1	0.20
13	1	0.25
14	2	0.35
15	1	0.40
16	3	0.55
17	2	0.65
18	2	0.75
19	1	0.80
20	1	0.85
21	1	0.90
22	1	0.95
23	0	0.95
24	1	1.00
25	0	1.00

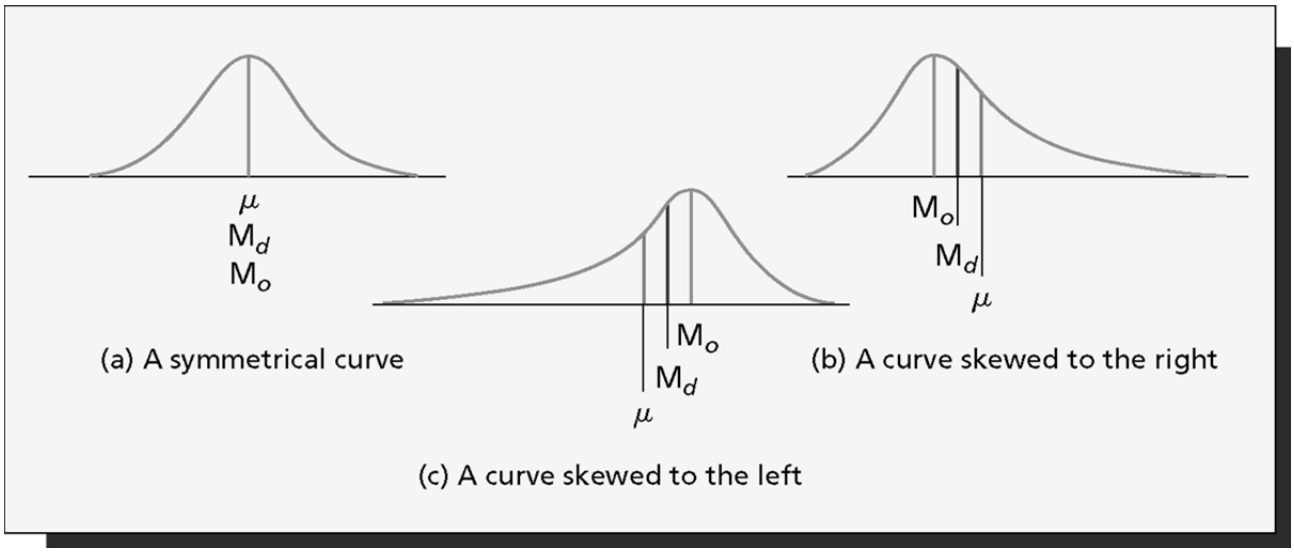


Sammenhængen mellem middeltallet, modus og medianen

Forholdet mellem **middeltallet**, her betegnet ved $\boldsymbol{\mu}$, modus, her betegnet ved $\mathbf{M_0}$ og **medianen**, her betegnet ved $\mathbf{M_d}$ kan sammenfattes i illustrationen på næste side. På denne baggrund kan der udledes information om fordelingens af datas udseende.

a) Symmetri: Her er $M_0 = M_d = \mu$
b) Højreskæv: Her er $M_0 < M_d < \mu$ (flest data mod venstre)
c) Venstreskæv: Her er $\mu < M_d < M_0$ (flest data mod højre)

Det ses, at medianen er det mest robuste og stabile lokationsmål, mens middelværdien er det mest varierende lokationsmål.



I de to eksempler ses følgende: For materialet om indkomster findes:

$$\mu = 15.85 < M_0 = 16 \text{ og } M_d = 16 \qquad \rightarrow \text{materialet er svagt venstreskævt}$$

For materialet om karakterer:

$$\mu = 6.45 < M_o = 7 \text{ og } M_d = 7 \qquad \rightarrow \text{materialet er svagt venstreskævt}$$

Afslutningsvis ses det, at fordelingen af befolkningen i Kina er højreskæv, når histogrammet er tegnet korrekt.

Kvartiler og percentiler

Udtrykket for medianen kan generaliseret til at finde andre punkter i en fordeling. Benævn en *kvartil* Q_i. En kvartil opdeler data i fjerdedele. Kvartilen defineres som:

$$\textit{Kvartilen} = q(n + 1) \textit{ ordnede position}$$

Her er $q = 0.25$ lig med den *nedre kvartil*, der kaldes Q_1. Hvis $q = 0.75$ fås den *øvre kvartil*, der kaldes Q_2. Hvis $q = 0.50$ fås medianen.

P*ercentilen* benævnes også *fraktilen* og defineres som:

$$\textit{Percentilen} = p(n + 1) \textit{ ordnede position}$$

Hvor p er et tal mellem 0 og 1. Deciler er et specialtilfælde af percentiler. Specielt gælder, at for $p = 0.10$ haves *1. decil*, mens for $p = 0.90$ fås *9. decil.*

Der kan beregnes kvartiler og deciler for datasættet om indkomster ved anvendelse af det netop givne formler. Man kan beregne **5-punktsopsummeringen** som:

1. decil	er 0.10-percentilen		= 9
Nedre kvartil	er 0.25-percentilen	(Q_1)	= 13
Medianen	er 0.50-percentilen		= 16
Øvre kvartil	er 0.75-percentilen	(Q_3)	= 18
9. decil	er 0.90-percentilen		= 19

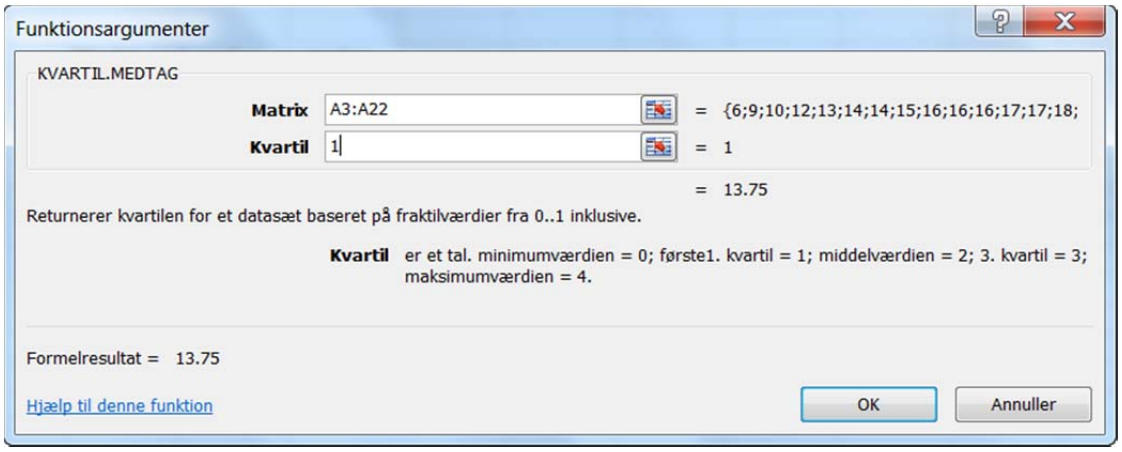
I tabellen nedenfor er vist, hvordan man kan aflæse værdierne.

Data	6	9	10	12	13	14	14	15	16	16	16	17	17	18	18	19	20	21	22	24
Frekvens	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05
Kumuleret	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95	1.00
Nummer	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*

Mere præcise værdier fås ved anvendelse af **Excel:**

10:	(20+1)0.10 = 2.10	Observation nr. 2.10 giver værdien	= 9.10
25:	(20+1)0.25 = 5.25	Observation nr. 5.25 giver værdien	= 13.75
50:	(20+1)0.50 = 10.50	Observation nr. 10.50 giver værdien	= 16.00
75:	(20+1)0.75 = 15.75	Observation nr. 15.75 giver værdien	= 18.25
90:	(20+1)0.90 = 18.90	Observation nr. 18.90 giver værdien	= 21.90

I **Excel** kan man finde *kvartilerne* ved anvendelse af udtrykket **Formler | Indsæt funktion | Statistisk | Kvartil.medtag**



Hvor: 0 = minimum, 1 = nedre kvartil, 2 = middelværdien, 3 = øvre kvartil og 4 = maksimum

Bemærk at værdierne er fundet mere præcist. Hertil er anvendt en metode, der kaldes lineær interpolation. Dette er *ikke* et krav til eksamen. Her er aflæsning tilstrækkeligt eller beregning med **lommeregneren.**

I **Excel** kan man finde *percentilerne* ved anvendelse af udtrykket **Formler | Indsæt funktion | Statistisk | Fraktil**

Hvor *K* angiver værdien af percentilen, der varierer mellem 0 og 1. I eksemplet er beregnet værdien for 1. decil i eksemplet med indkomsterne.

Der vendes tilbage til de grupperede datasæt i afsnit 8 og 9.

Geometrisk middeltal:

Det *geometriske middeltal* fås med at multiplicere alle observationerne med hinanden, og derefter at uddrage den *n´*te rod.

Mens middeltallet er additivt er det geometriske middeltal multiplikativt. Hvilket middeltal, der giver det mest retvisende billede af det materiale, der analyseres, afhænger af, hvilken underliggende data genererende proces (DGP), som antages at være den korrekte. Langt de fleste økonomiske variabler antages at være relaterede additivt, mens variabler, der relateret sig til eksempelvis evolution, ofte er multiplikativt relaterede. Inden for økonomi kan der forekomme multiplikativt relaterede processer eksempelvis relateret til hyperinflation eller variationen af visse typer af finansielle aktiver.

Det geometriske middeltal er defineret som:

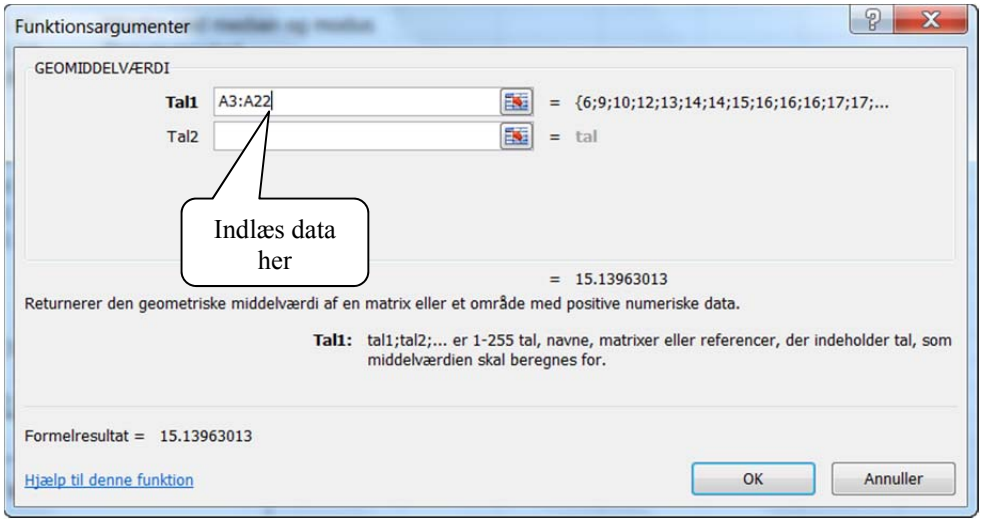
$$\overline{X}_G = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Det geometriske middeltal er *altid* mindre end det additive middeltal. I eksemplet med indkomsterne findes det geometriske middeltal som:

$$\overline{X}_G = \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[20]{6\times 9\times 10\times \ldots \times 21\times 22\times 24} = 15.14$$

Forfatteren til disse noter kan ikke finde en funktion, der kan beregne det geometriske middeltal, på lommeregneren.

Ved anvendelse af **Excel** findes det geometriske middeltal som **Fomler │Indsæt funktion │statistisk │Geomiddelværdi** som vist:



6. Spredningsmål og boksdiagrammet

Spredningsmål giver et udtryk for usikkerheden på det oftest observerede. Årsagen, til at der anvendes forskellige spredningsmål, er som i det forrige afsnit, at de forskellige mål inddrager forskellige mængder af information. Nogle af spredningsmålene skulle gerne være kendt fra folkeskolen eller gymnasiet.

I det følgende omtales:

- Variationsbredden, kvartilafstanden og boksdiagrammet
- Variansen og standardafvigelsen
- Variationskoefficienten
- Skævhed og topstejlhed (kurtosis)

Variationsbredden, kvartilafstanden og boksdiagrammet

Fælles for disse spredningsmål er, at ikke al information i datasættet anvendes.

Variationsbredden er lig med maksimum minus minimum. For datasættet om indkomster er variationsbredden lig med 24 − 6 = 18, mens det for datasættet med fordelingen af karakterer er lig med 12 – 2 = 10.

Kvartilafstanden er lig med den øvre kvartil minus den nedre kvartil. Inden for denne afstand forventes 50 procent af data at være lokaliseret. Kvartilafstanden forkortes IQR = $Q_3 - Q_1$ for Inter Quartile Range. I eksemplet med indkomsterne fås kvartilafstanden til at være lig med IQR = $Q_3 - Q_1$ = 18.75 − 13.25 = 4.50.

Decilafstanden er lig med den 9. decil minus den 1. decil. Inden for denne afstand forventes 80 procent af data at være lokaliseret. I eksemplet med indkomsterne fås decilafstanden til at være lig med 21.90 − 9.10 = 12.80.

Et *boksdiagram* anvender information om median og kvartiler til at lokalisere *ekstreme observationer*. Disse kan ofte påvirke hele fordelinger og bør således behandles specielt. De ekstreme observationer opdeles i to kategorier dels *outliers* dels *suspected outliers.* De to typer af ekstremer identificeres som følger:

- En *outlier* findes, hvis observationen ligger mere end 3 gange kvartilafstanden fra den nedre eller øvre kvartil
- En *suspected outlier* findes, hvis observationen ligger mere end 1.5, men mindre end 3 gange kvartilafstanden fra den nedre eller øvre kvartil

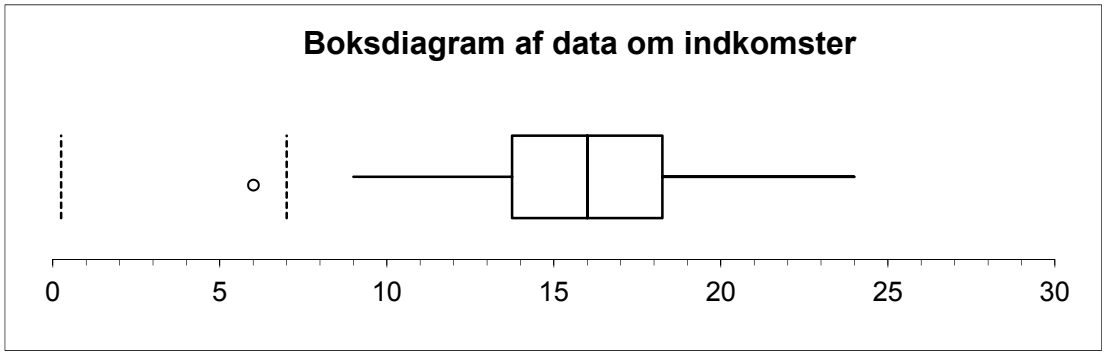
Boksdiagrammet finder meget hyppigt anvendelse inden for den deskriptive statistik, og giver et hurtigt visuelt overblik af fordelingen af et datasæt. Boksdiagrammet har to svagheder. For det første anvendes ikke alle observationer til konstruktionen af diagrammet, og for det andet er valget af de 3 værdier 1.5 og 3 ikke umiddelbart gennemskueligt.

Man kan konstruere et boksdiagram for datasættet med indkomsterne ved indledningsvis at udregne nedre og øvre grænser for *outliers* og *suspected outliers.* Disse betegnes som henholdsvis *inner* og *outer fence.*

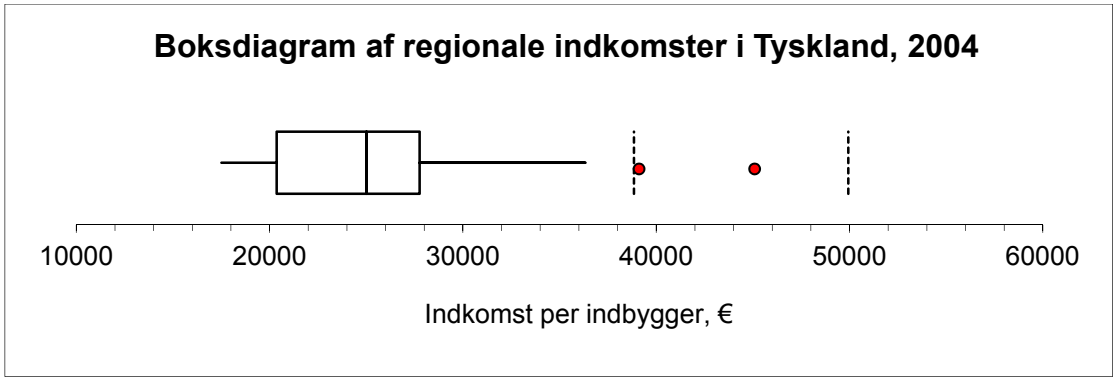
Nedre inner fence: $Q_1 - 1.5 \times IQR = 13.75 - 1.5(4.5) = 7.00$
Nedre outer fence: $Q_1 - 3.0 \times IQR = 13.75 - 3.0(4.5) = 0.25$

Øvre inner fence: $Q_3 + 1.5 \times IQR = 18.25 + 1.5(4.5) = 25.50$
Øvre outer fence: $Q_3 + 3.0 \times IQR = 18.25 + 3.0(4.5) = 32.25$

Boksdiagrammet kan nu opstilles som følger:



Her er anvendt et statistisk program, der hedder Megastat til at tegne diagrammet, men det er let at lave manuelt. Det ses, at den mindste værdi ”6” i datasættet er en *suspected outlier.* Datasættet om indkomster er således rimeligt homogent.



Betragt som et andet eksempel ovenstående boksdiagram, der viser indkomsten per indbygger i € for regioner i Tyskland for 2004. Det kan ses, at der er 2 *suspected outliers*. Det er Hamburg med en indkomst per indbygger på 45,00 € og München (Oberbayern) med en indkomst på 39,000 €. Den laveste indkomst på godt 17,500 € findes i region Dessau.

Variansen og standardafvigelsen

Variansen og standardafvigelsen er de oftest anvendte mål for spredningen i et datasæt. Standardafvigelsen er kvadratroden af variansen.

Ganske som det er tilfældet med middeltallet, så medgår alle observationerne til at beregne variansen. Variansen beregner summen af de kvadrerede afvigelser fra en given observation til middelværdien. Afvigelsen er kvadreret, da der findes observationer såvel under som over middelværdien. Blev der ikke taget et kvadrat ville de positive og negative afvigelser i forhold til middelværdien gå ud mod hinanden, og variationen ville forsvinde.

For et datasæt i en stikprøve er standardafvigelsen givet som:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Her angiver x_i en given observation i datasættet, mens n er antallet af observationer. Der divideres med $n-1$, da s betegner standardafvigelsen i en stikprøve. Ved at dividere med $n-1$ fås en lidt større spredning end i en totalpopulation.

For en stikprøve i en totalpopulation er standardafvigelsen givet som:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

Det ses, at for stigende størrelse af stikprøven bliver forskellen mellem s og σ meget lille. **Generelt anvendes altid målet for variansen eller standardafvigelsen i stikprøven.** På lommeregneren fås begge mål.

En hyppigt anvendt formel for standardafvigelsen, som dog er vanskeligere at fortolke, er givet som:

$$s = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]}$$

For en *grupperet fordeling* med k grupper med frekvensen f_i haves:

$$s = \sqrt{\frac{\sum_{i=1}^{k} f_i \times (x_i - \bar{x})^2}{n-1}}$$

Betragt som eksempel standardafvigelsen for datasættet om indkomster findes som:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{(6-15.85)^2 + (9-15.85)^2 + ... + (22-15.85)^2 + (24-15.85)^2}{20-1}} = 4.46$$

Med den anden formel for standardafvigelsen fås:

$$s = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]} = \sqrt{\frac{1}{20-1}\left[5{,}403 - \frac{(317)^2}{20}\right]} = \sqrt{\frac{1}{19}\left[5{,}403 - 5{,}024.45\right]} = \sqrt{19.92} = 4.46$$

Som forventet. De to sumstørrelser kan findes på **lommeregneren**. Mere herom i næste afsnit.

Betragt, som et eksempel på en grupperet fordeling, datasættet fra tidligere, der viser resultaterne af bestående studenter, fra eksamen i International Økonomi fra BA int. i Flensburg fra februar 2011. Her fås:

$$s = \sqrt{\frac{\sum_{i=1}^{k} f_i (x_i - \bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{10 \times (2-6.45)^2 + 26 \times (4-6.45)^2 + 33 \times (7-6.45)^2 + 19 \times (10-6.45)^2 + 4 \times (12-6.45)^2}{92-1}} = 2.83$$

Man kan diskutere om dette er en stikprøve. Divideres med $n = 92$ fås at $\sigma = 2.81$.

Variationskoefficienten
Variationskoefficient er et udtryk for den relative variation i et datasæt. Den betegnes med *CV* og beregnes som standardafvigelsen divideret med middeltallet:

$$CV = \frac{s}{\overline{X}}$$

Variationskoefficienten er velegnet til sammenligning mellem forskellige datasæt.

- Hvis datasættet har en stor variation, så vil *CV* være stor
- Hvis datasættet har en lille variation, så vil *CV* være lille

Skævhed og topstejlhed (kurtosis)
Der kan beregnes matematiske udtryk for en fordelingens afvigelse fra middelværdien kaldet *skævheden* og dens koncentration. Sidstnævnte betegnes også topstejlheden eller *kurtosis*.

Skævheden betegnes *SK*. Den beregnes matematisk som:

$$SK = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3 / n}{s^3} \qquad i = 1,2,...,n$$

Hvis SK > 0 er fordelingen højreskæv, mens fordelingen er venstreskæv hvis SK < 0. Hvis SK = 0 er fordelingen symmetrisk. Se også illustrationen på side 18.

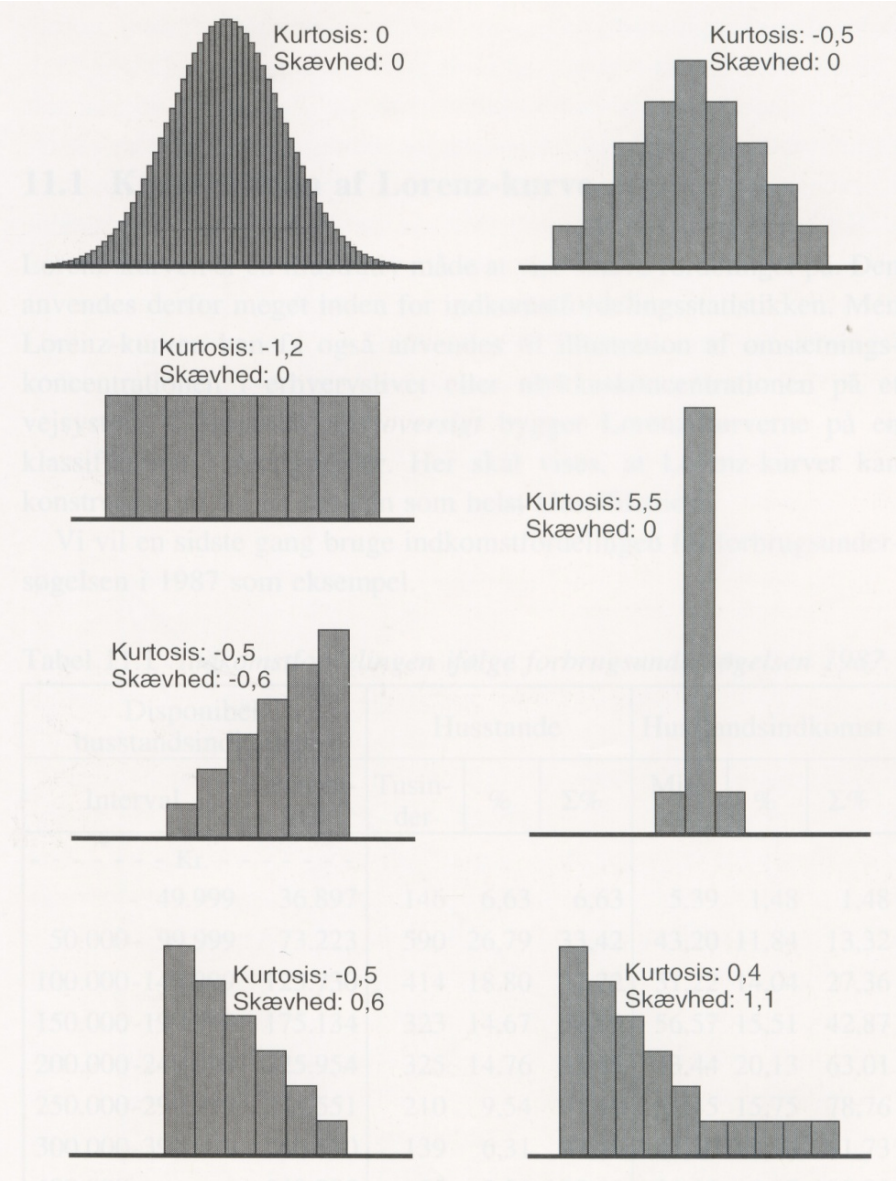
Topstejlheden eller kurtosis beregnes som:

$$KU = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4 / n}{s^4} \qquad i = 1,2,...,n$$

Hvis KU = 0 haves Normalfordelingen. Dette kaldes også en *mesokurtisk* (ideal) fordeling Hvis KU > 0 haves en *leptokurtisk* (stejl) fordeling, og hvis KU < 0 haves en *platykurtisk* (flad) fordeling. Egentlig kan KU ikke være negativ, men i mange programmer fratrækkes værdien 3. Dette er af tekniske årsager.

Der forventes ikke at blive stillet eksamensspørgsmål i beregning af SK og KU, men måske i fortolkningen.

På den følgende side er gengivet en række eksempler på værdier af SK og KU taget fra E.M. Bøye, 2003, *Deskriptiv Statistik*, 3. udgave, forlaget Swismark, side 205.

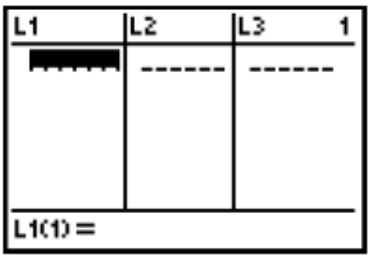


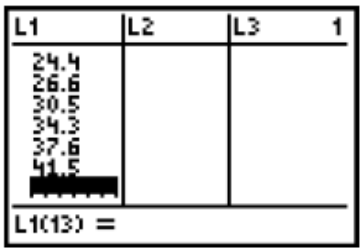
7. Deskriptiv statistik på lommeregneren og i Excel

Næsten alle de ovenfor anførte positions- og spredningsmål kan udføres på **lommeregneren.** Dette er helt i orden til eksamen, og der kan, som nævnt tidligere, spares meget tid ved være fortrolig med funktionerne på lommeregneren. Man kan endda tegne et boksdiagram! Sidstnævnte kan dog ikke anbefales til eksamen. Her er det meget hurtigere at tegne på papir!

Prøv også at læse i manualen til lommeregneren. Den findes dels på den medfølgende CD, dels kan den hentes på Internettet på adressen www.education.ti.com/danmark

På **TI-84eren** skal man indledningsvis indlæse tallene. Der er 6 registre til at lagre data i kaldet L_1 til L_6. Tryk STAT→1: EDIT→ENTER, så kommer menuen frem til venstre på næste side. Data hulles ind som vist til højre





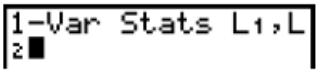
Der kan laves mange forskellige manipulationer med data i registre – se manualen!

Næste trin er at beregne positions- og spredningsmål. For et **normalt datasæt** vælges: STAT→CALC vælg nu 1: **1-var stats** ENTER→**L_1**→ENTER. I dette eksempel er findes data i registret L_1.

For et **grupperet datasæt** skal man indtaste både variablen samt frekvensen. Det vil sige, at man skal indtaste data i to registre. I eksemplet med karakterne fra faget International Økonomi kunne L_1 indeholde de beståede karakterer dvs. 2,4,7,10 og 12, mens L_2 kunne indeholde frekvenserne.

Nu vælges på lommeregneren: STAT→CALC vælg nu 1: **1-var stats** ENTER→**L_1,L_2**→ENTER

PS: "Komma" findes lige over ciffertesterne "7 til 9". Se illustration nedenfor til venstre.



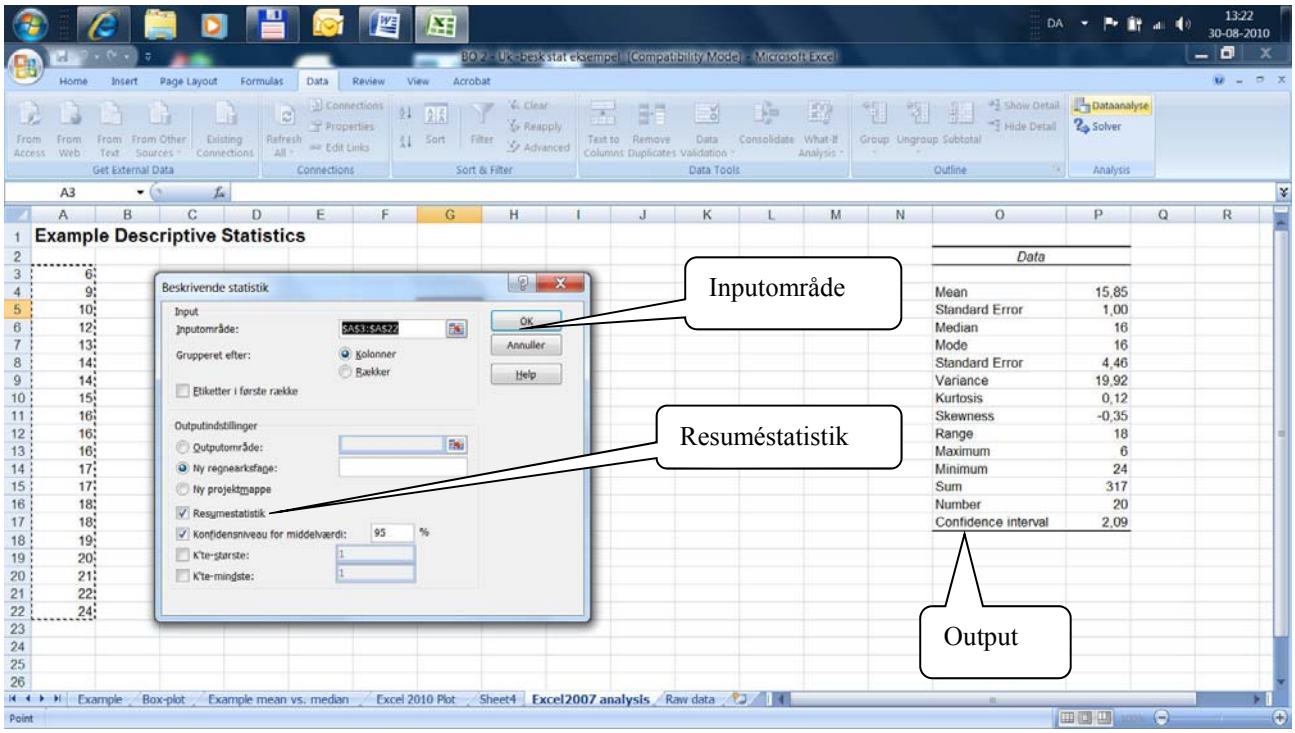
Der kan også laves et **boksdiagram.** Antag man ønsker at lave et diagram for data om indkomsterne som det, der er udarbejdet på side 23. Lad data være indlæst i registret L_1,

Det vides at minimum er lig med 6 og maksimum er lig med 24. Det vil sige at en bredde på 30 enheder er passende. Nu vælg 2ND→STAT PLOT→1:ENTER (testen "STAT PLOT findes til venstre lige under lommeregnerens display).

Se også illustrationen til højre ovenfor. Plottet skal være "ON" og figurtypen er "Boksdiagram". Xlist is L_1 og tryk nu WINDOW og sæt Xmin=0, Xmax=30 og så tryk ZOOM vælg 9: ZoomStat ENTER. Nu kommer boksdiagrammet (forhåbentlig) på skærmen.

Deskriptiv statistik ved anvendelse af Excel

De centrale positions- og spredningsmål kan let beregnes ved anvendelse af **Excel.** Dette gøres som følger ved i DATA at vælge DATA ANALYSIS, og så vælge BESKRIVENDE STATISTIK. Dette er gjort på skærmbilledet nedenfor.



Udskriften er vist i skærmbilledet til højre og nedenfor i forfatterens oversættelse. Det kan valgfrit vælges, om man vil medtage et konfidensinterval.

Beskrivende statistik for indkomster		
Middeltallet	15.85	Mindre end median og modus
Standardfejl	0.998	s/(rod n) – ikke gennemgået
Median	16	
Modus / Typetal	16	
Standardafvigelse	4.46	
Stikprøvevarians	19.92	
Kurtosis	0.12	Lidt stejl fordeling
Skævhed	-0.35	Venstreskæv
Variationsbredde	18	
Minimum	6	
Maksimum	24	
Sum	317	
Antal	20	

Konfidensinterval (95,0%)	2.09	Omtales i notesæt 3

8. Grupperede datasæt: Lokation, spredning og boksdiagram

I dette afsnit ses der på et mere kompliceret datasæt for indkomster for Danmark. Materialet er fordelt på intervaller og for hver interval er den gennemsnitlige indkomst beregnet. Havde denne indkomst ikke været tilgængelig havde man i stedet skulle have anvendt midtpunktet for indkomstintervallet.

Hvordan beregnes nu den samlede gennemsnitlige indkomst og standardafvigelsen? Her anvendes de tidligere viste formler. Bemærk at der er tale om en totalpopulation. Der opstilles et regneark af form som nedenstående tabel.

Disponible husstandsindkomster, Danmark, 1987

	Interval for ansættelse af indkomst 1,000 DKK	Antal hustande, 1,000	Gns. indkomst 1,000 DKK	Indkomst-Masse Mio. DKK	Afvigelse	Kvadrat	
i		f_i	x_i	$f_i \times x_i$	$(x_i - \mu)$	$(x_i - \mu)^2$	$f_i \times (x_i - \mu)^2$
1	0 - 49.9	146	36.9	5,387	-128.7	16563.69	2418298
2	50 - 99.9	590	73.2	43,202	-92.4	8537.76	5036983
3	100 - 149.9	414	123.7	51,224	-41.9	1755.61	726822
4	150 - 199.9	323	175.1	56,568	9.5	90.25	29151
5	200 - 249.9	325	225.9	73,435	60.3	3636.09	1181729
6	250 299.9	210	273.6	57,446	108.0	11664.00	2449440
7	300 399.9	139	340.6	47,339	175.0	30625.00	4256875
8	400 -	55	548.3	30,156	382.7	146459.29	8055261
Sum		2,202		364,757			24154559

Kilde: Danmarks Statistik, Statistisk Årbog, 1994, side 220-221.

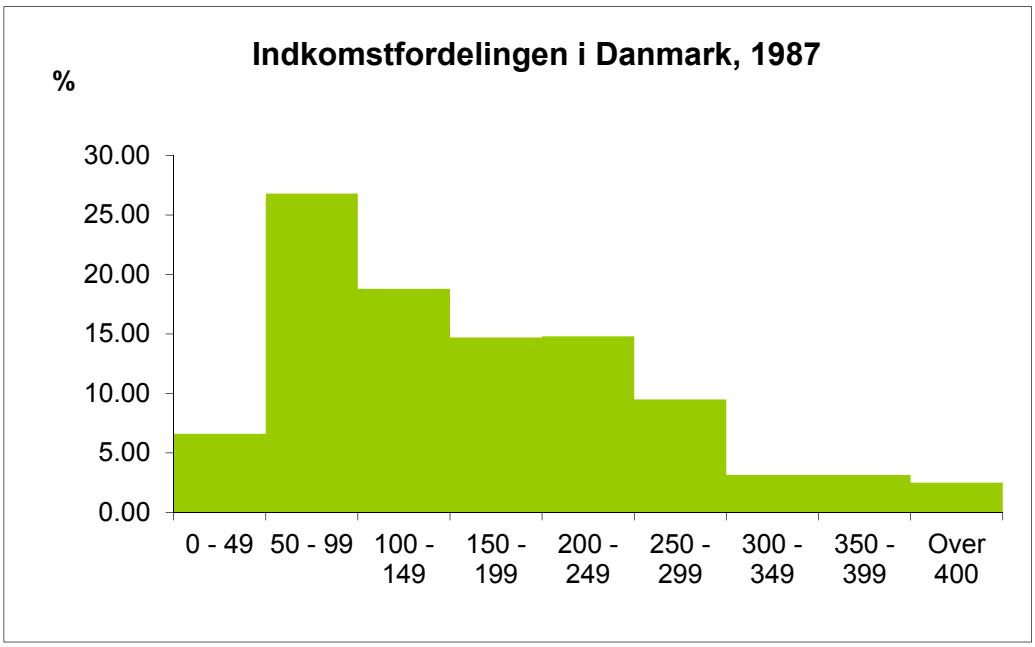
Der er 8 grupper, så $k = 8$. Nu indsættes der i formlerne:

Middeltal:
$$\mu = \frac{\sum_{i=1}^{k} f_i \times x_i}{n} = \frac{364{,}757}{2{,}202} = 165{,}648\ DKK \approx 165.6$$

Standardafvigelsen:
$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} f_i \times (x_i - \mu)^2}{n}} = \sqrt{\frac{24{,}154{,}559}{2{,}202}} = 104.73$$

Der er således en pæn spredning af indkomsterne, hvilket også fremgår af det histogram, der findes på næste side, som viser den procentuelle fordeling af indkomsterne.

Hvordan beregnes nu medianen og kvartiler præcist for dette grupperede datasæt, hvor materialet optræder som intervaller? Hertil kan man anvende *lineær interpolation*. Det er en metode, hvor man inden for intervallet beregner, hvor stor en andel af materialet, der ligger i en given kategori.



For at kunne beregne medianen med mere foretages en supplerende beregning på materialet i tabellen ovenfor. Først beregnes den relative frekvens idet n = 2,202. Fra den relative frekvens kan man beregne den kumulative frekvens.

Disponible husstandsindkomster, Danmark, 1987

i	Interval for ansættelse af indkomst 1,000 DKK	Antal hustande, 1,000 f_i	Antal hustande frekvens, % f_i/n	Kumuleret frekvens, %
1	0 - 49.9	146	6.6	6.6
2	50 - 99.9	590	26.8	33.4
3	100 - 149.9	414	18.8	52.2
4	150 - 199.9	323	14.7	66.9
5	200 - 249.9	325	14.8	81.7
6	250 299.9	210	9.5	91.2
7	300 399.9	139	6.3	97.5
8	400 -	55	2.5	100.0
Sum		2,202	100.0	

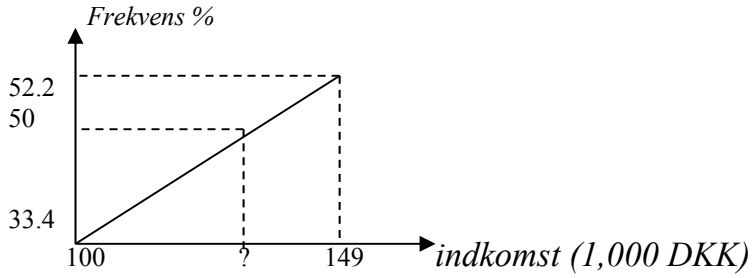
Kilde: Danmarks Statistik, Statistisk Årbog, 1994, side 220-221.

Umiddelbart kan man se fra kolonnen med den kumulative frekvens, at den nedre kvartil vil befinde sig i intervallet mellem 50 og 99.9 tusinde DKK. Den nedre kvartil svarer til en andel på 25 procent. Tilsvarende for medianen og den øvre kvartil.

For at beregne dette præcist anvendes som nævnt *lineær interpolation.* Hertil anvendes følgende formel, som forklares på næste side:

$$Værdi = \text{“Enden af intervallet”} - \frac{\text{"For langt relativt til værdi"}}{\text{"Total bredde i procent pct"}} \times \text{intervalbredden}$$

Med de tal, som er givet i tabellen nedenfor, kan der for eksempelvis beregningen af *medianen* opstilles en illustration som følger:



Hvad sker der? Medianen findes ved den observation, der findes i midten af datasættet. Det vil sige ved en fraktil på 50 procent. Af tabellen ovenfor ses, at 52.2 procent af alle indkomster findes ved en indkomst på 149,999 DKK eller mindre, mens 33.4 procent af indkomsterne findes ved en indkomst på 100,000 DKK eller mindre. Medianen vil således befinde indenfor dette interval.

Antager man, at alle indkomsterne er jævnt fordelt inden for intervallet, kan man finde den indkomst, der vil være gældende ved 50 procent. I forhold til de 50 procent er man ”2.2 procent for langt” ved en indkomst på 149.999 DKK. Man skal således gå ”2.2 procent tilbage” i forhold til bredden på hele intervallet. Denne bredde er lig med 52.2−33.4 = 18.8 procent. Bredden af dette interval ses af den tredje kolonne i tabellen ovenfor.

Brug af formlen ovenfor giver da:

Medianen: $$150{,}000 - \frac{(52.2 - 50)}{18.8} \times 50{,}000 = 150{,}000 - 5{,}851 = 144{,}149$$

Tilsvarende fås for kvartiler og deciler:

Nedre kvartil: $$100{,}000 - \frac{(33.4 - 25)}{26.8} \times 50{,}000 = 84{,}328 \qquad (Q_1)$$

Øvre kvartil: $$250{,}000 - \frac{(81.7 - 75)}{14.8} \times 50{,}000 = 227{,}365 \qquad (Q_3)$$

Første decil: $$100{,}000 - \frac{(33.4 - 10)}{26.8} \times 50{,}000 = 56{,}343$$

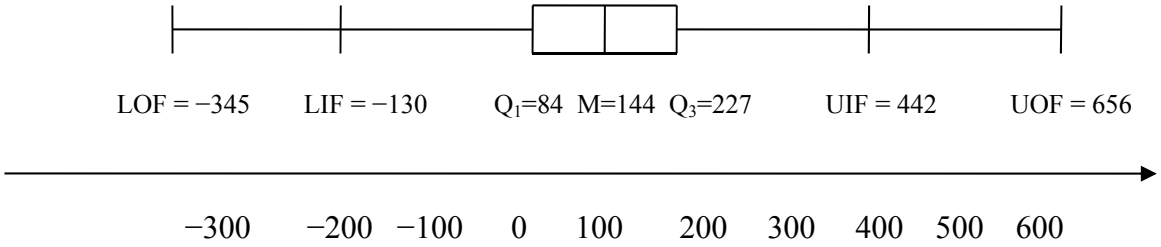
Niende decil: $300{,}000 - \frac{(91.2 - 90)}{9.5} \times 50{,}000 = 293{,}684$

Kvartilafstanden: *(IQR):* $(Q_3 - Q_1) = 227{,}365 - 84{,}328 = 143{,}037$

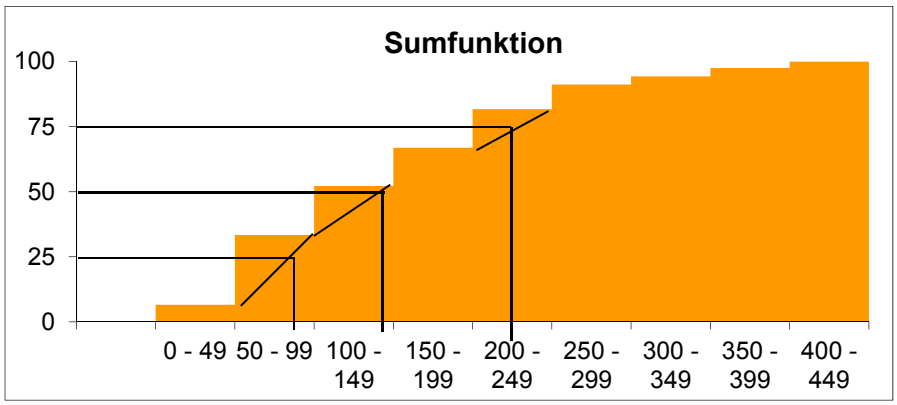
Som tidligere i disse noter, kan man beregne for boksdiagrammet:

Nedre inner fence:	$Q_1 - 1.5 \times IQR = 84{,}328 - 1.5(143{,}037) =$	−130,228
Nedre outer fence:	$Q_1 - 3.0 \times IQR = 84{,}328 - 3.0(143{,}037) =$	−344,783
Øvre inner fence:	$Q_3 + 1.5 \times IQR = 227{,}365 + 1.5(143{,}037) =$	441,921
Øvre outer fence:	$Q_3 + 3.0 \times IQR = 227{,}365 + 3.0(143{,}037) =$	656,476

Boksdiagrammet ser nu ud som følger:



Afslutningsvis kan en *sumfunktion* tegnes som følger, idet der er indsat median og kvartiler:

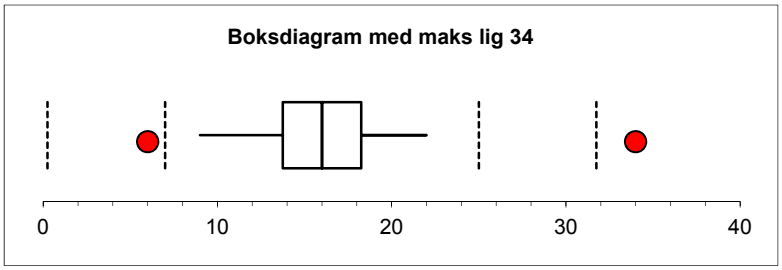


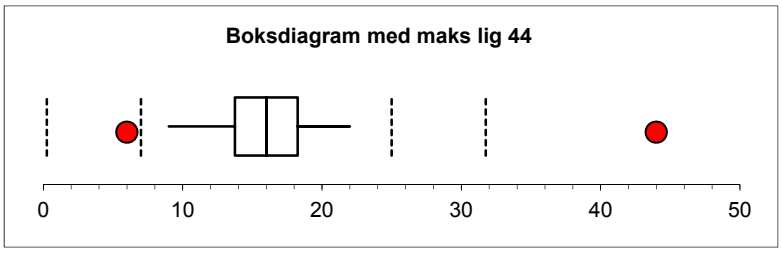
9. Deskriptiv statistik: Identifikation af ekstremer

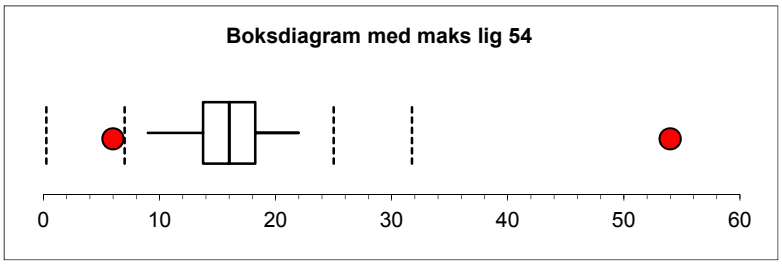
Hvad er konsekvensen af ekstremer? For at undersøge denne problemstilling foretages et lille eksperiment. Med udgangspunkt i eksemplet om indkomster, øges den maksimale værdi fra 24 i udgangssituationen til henholdsvis ”34”, ”44” og ”54”. Resultatet vises nedenfor dels som ”beskrivende statistik” dels som boksdiagrammer.

Det ses, at mens middeltallet øges, så forbliver medianen konstant. Det er netop derfor, at medianen er god til eksempelvis statistik om lønninger. Bemærk også dels at fordelingen bliver mere og mere højreskæv, dels at standardafvigelsen øges kraftigt. En enkelt ekstrem observation kan således have stor indflydelse på et datasæt!

	Basis	*Maks=34*	*Maks=44*	*Maks=54*	
Middeltallet	15.85	16.35	16.85	17.35	Stiger
Standardfejlen	0.99	1.29	1.69	2.13	
Medianen	16	16	16	16	Konstant!
Modus / Typetallet	16	16	16	16	
Standardafvigelsen	4.46	5.79	7.56	9.52	
Stikprøvevariansen	19.92	33.50	57.08	90.66	
Kurtosis	0.12	3.88	8.99	12.55	
Skævhed	-0.35	1.19	2.43	3.16	Stiger
Variationsbredden	18	28	38	48	
Minimum	6	6	6	6	
Maksimum	24	34	44	54	
Sum	317	327	337	347	
Antal	20	20	20	20	
Konfidensinterval (95.0%)	2.09	2.71	3.54	4.46	Stiger







Nils Karl Sørensen ©
Statistik I, BA int Flensborg

Sæt 2: Sandsynlighedsteori og statistiske fordelinger

af Nils Karl Sørensen

Til afsnittene om fordelinger kan sandsynlighederne enten udregnes ved anvendelse af lommeregneren, eller der kan anvendes et statistisk tabelværk. Sidstnævnte kaldes her for **Statistics Tables**. Tabellerne er beregnet ved anvendelse af regnearket Excel. Tabelværket kan findes i Blackboard under *Statistiske Tabeller*. De statistiske tabeller skal også anvendes i Statistik II til foråret.

Indhold *side*

1. Udfald og hændelser

Fætter Højben fra Anders And er en ufattelig heldig person! Måske vinder han lotteriet i Andeby ikke én gang, men *to* gange. Hvor sandsynligt er dette?

For at kunne undersøge denne problemstilling, må der fastlægges begreber. Først defineres begreberne *sandsynlighed* og *udfald.*

Udfaldet afhænger af dels af lotteriet dels af antallet af deltagere. Antallet af deltagere må være beboere i Andeby, som deltager i lotteriet. Lotteriet kan være en *simpel tilfældig udvælgelse* af vinderen. Deltageren i spillet kan skrive sit navn på en seddel, som lægges i en sort boks. Boksen rystes, når alle deltagerne har lagt deres seddel i boksen, og en person med bind for øjnene trækker en seddel med navnet på vinderen. Dernæst gentages hele processen. I begge tilfælde vinder Fætter Højben. Der er således tale om *simpel tilfældig udvælgelse med tilbagelægning.*

Man kan definere en sandsynlighed som følger:

Sandsynlighed = kvantitativt mål for usikkerhed

Der er således tilknyttet en numerisk værdi til sandsynligheden. Sandsynligheden er baseret på et *eksperiment.* I eksempelet ovenfor er selve trækningen af lotteriet eksperimentet, mens vinderen af eksperimentet er udfaldet af eksperimentet.

For en *sandsynlighed* skal der være opfyldt to betingelser:

1. En sandsynlighed, som er tilordnet et eksperimentalt udfald, har et udfald mellem 0 og 1. Hvis *E* betegner udfaldet, og *P(E)* betegner sandsynligheden, vil der gælde at $0 \le P(E) \le 1$.
2. Summen af alle eksperimentale udfald er lig med 1.

Betragt som eksempel et eksperiment, hvor der kastes med en mønt. Der er to udfald: Krone (H=head) og plat (T=tail). Hvis mønten er fair og uden mærker, så vil udfaldet af eksperimentet ”kast af mønten” være lig med 0.5 for at få enten ”krone” eller ”plat”. Summen af alle udfaldene er lig med 0.5 + 0.5 = 1. Det fremgår, at begge betingelserne er opfyldt.

Hvordan kan vi være sikre på, at mønten er fair? Vi kan gentage eksperimentet ”kast af mønten” eksempelvis 250 gange. Hvis mønten er fair, så vil der være omkring 125 ”krone” og 125 ”plat”. Man kan beregne sandsynligheden for ”krone” af eksperimentet som følger:

$$P(H) = \frac{antal\ udfald\ af\ krone\ (gunstige\ udfald)}{antal\ af\ eksperimenter\ (mulige\ udfald)} \approx 0.5$$

Det antages, at antallet af eksperimenter kan øges uendeligt. Dette udsagn kaldes *Cordanos regel*. Desto flere forsøg desto mere præcis bliver beregningen af sandsynligheden.

Den beregnede sandsynlighed for eksemplet med kast med mønten er baseret på det observerede udfald af et eksperiment. Dette kaldes også den *objektive sandsynlighed.* En sandsynlighed kan blive tilordnet en hændelse på to forskellige måde. Enten som en *objektiv sandsynlighed* eller som en *subjektiv* sandsynlighed:

1. *Objektiv sandsynlighed:* Er baseret på beregnede værdier af observerede hændelser. Dette kaldes også for den *klassiske sandsynlighed*.

2. *Subjektiv sandsynlighed:* Er baseret på formodninger, information, intuition og andre subjektive kriterier. Anvendelsen af denne type af sandsynlighed er af nyere dato, og blev udviklet i 1930'erne. Begrebet er kontroversielt. Subjektiv sandsynlighed benævnes også *personlig sandsynlighed.* Betragt som et eksempel en vejrmelding. Her kan analytikeren anvende information fra forskellige modelsimulationer af udvikling i vejret. Derfra vælges det mest sandsynlige forløb ud fra personlige (subjektive) overvejelser.

Uanset hvilken type for sandsynlighedsbegreb der anvendes, gælder de samme matematiske regler for regnerier og analyser.

For at uddybe sandsynlighedsbegrebet anvendes nogle begreber for mængdelæren fra matematik i folkeskolen.

Definer:

- En *mængde* er en samling af elementer, hvor elementer er alt, der kan tilordnes en nummer/værdi.

- Den *tomme mængde* indeholder ingen elementer og benævnes "**Ø**"

- Den *universelle mængde* indeholder alle elementer og benævnes **S**

Betragt nu mængden af elementer **A.** Det kan eksempelvis være alle personer, der deltager i lotteriet i Andeby. Sandsynligheden *P(A)* for at vinde i lotteriet kan skrives som:

$$0 \leq P(A) \leq 1$$

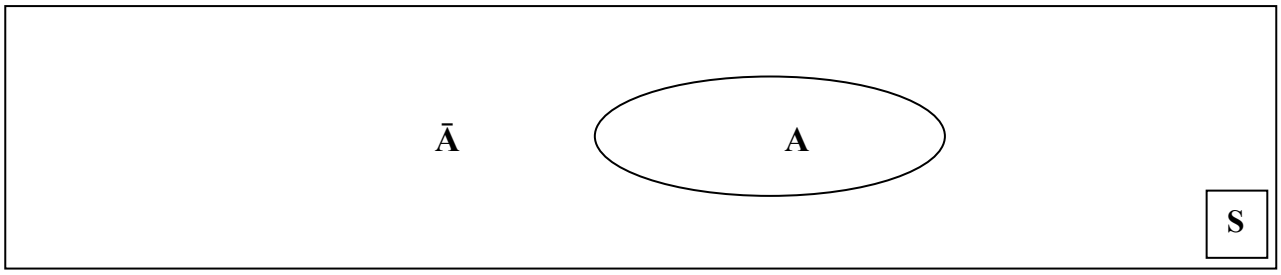
Givet mængden **A** kan man definere den **komplementære** mængde kaldet $\bar{\mathbf{A}}$, som alle de, der *ikke* deltager i lotteriet i Andeby. Dette skrives som:

- Den *komplementære mængde* til **A** er mængden af alle elementer i **S**, der ikke er i **A**

Det vil sige, at der gælder, at sandsynligheden for den komplementære hændelse kan skrives som $P(\overline{A}) = 1 - P(A)$

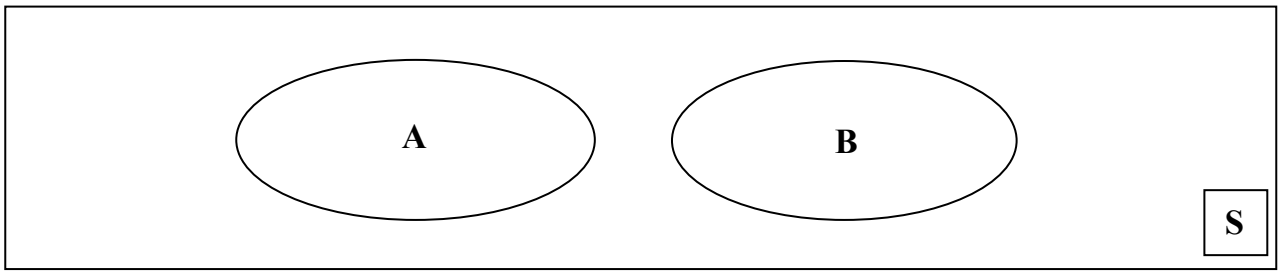
Man kan anvende begreberne til at opstille *Venn-diagrammer.* Formålet med at opstille et Venn-diagram er, at lette beregningen af sandsynligheder ved at øge overskueligheden.

Et *Venn-diagram* for den komplementære hændelse kan opstilles som:



Rammen angiver mængden af alle udfald **S.** Denne placeres normalt i nederste højre hjørne. I dette tilfælde er **S** lig med summen af de to hændelser. Det vil sige alle i Andeby. **A** er alle elementer af udfald i mængden som spiller, mens **Ā** er resten. Det vil sige det komplementære udfald.

Vi introducerer nu en ekstra mængde af hændelser kaldet **B.** Mængden **B** er alle personer i Gåserød, som deltager i lotteriet i denne by, som ligger lidt væk fra Andeby. Antag at lotteriet i Gåserød kun er for beboerne i denne by, og det samme er tilfældet for lotteriet i Andeby. Da er de to hændelser gensidigt udelukkende eller *uafhængige*. *Venn-diagrammet* for denne situation er angivet som følger:

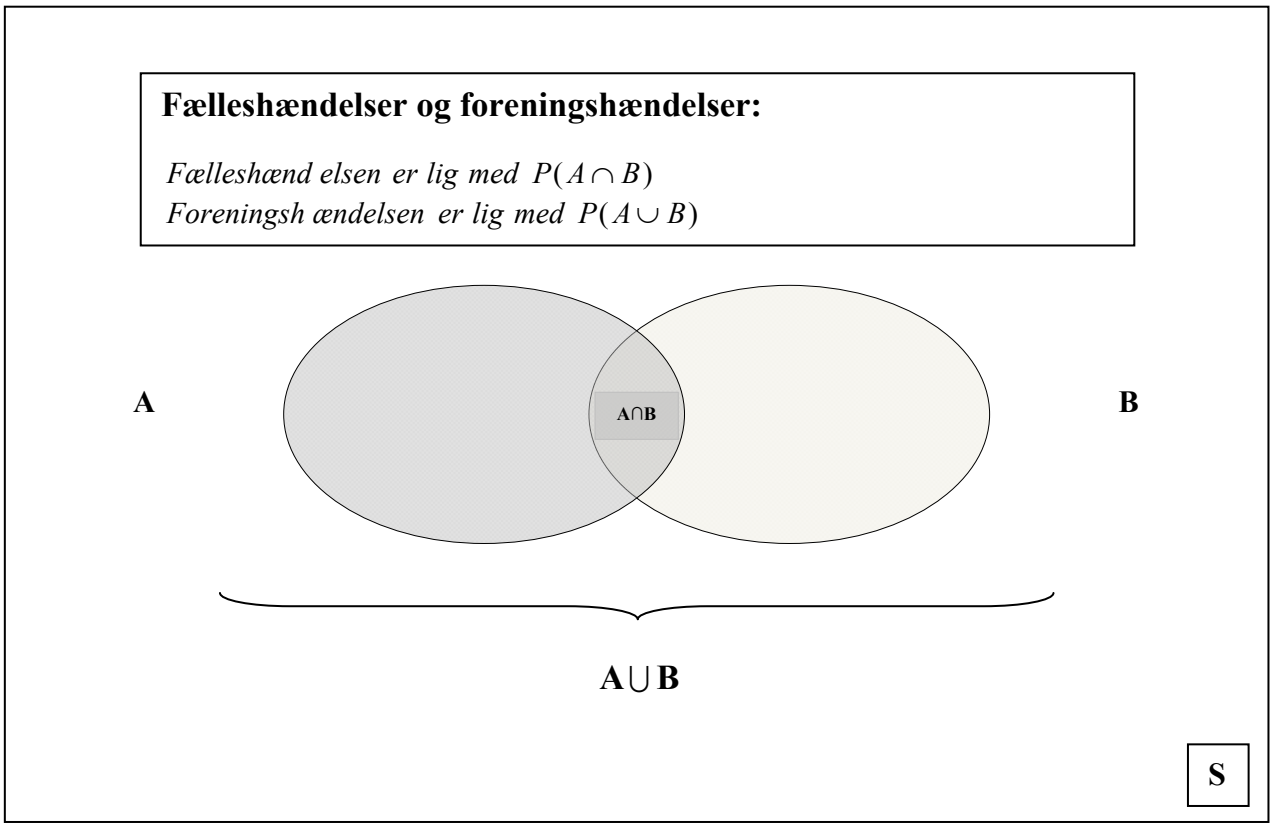


Her er **A** og **B** personerne i de to byer, som deltager i lotterierne, mens **S** er alle beboerne i de to byer tilsammen.

Lad os nu tillade, at deltagerne i lotteriet i Andeby også må deltage i lotteriet i Gåserød og omvendt. Dette komplicerer situationen. For at kunne håndtere situationen introduceres begreberne *foreningshændelser* og *fælleshændelser.*

Foreningshændelsen	= Består af de udfald, der indgår i både **A** og **B**
Fælleshændelsen	= Består af de udfald, der indgår i enten **A** eller **B**

Følgende *Venn-diagram* anskueliggør sammenhængen:



Diagrammet kan anvendes til at vise nogle regler for sandsynligheder. For de gensidigt udelukkende hændelser gælder, at fælleshændelsen er lig med nul. Det vil sige, at $P(A \cap B) = 0$. *Foreningshændelsen* findes to ved summation af de to mængder. Det vil sige:

$$P(A \cup B) = P(A) + P(B)$$

Dette kaldes også *additionsreglen.* Hvad nu, hvis dette ikke er tilfældet? I eksemplet kunne man ophæve restriktionen om, at borgerne kun må deltage i lotteriet i den by, som de er indbyggere i. I dette tilfælde vil der gælde for *fælleshændelsen* at $P(A \cap B) \neq 0$. *Foreningshændelsen* findes da som:

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

Hvad sker der her? Der er både nogle fra Andeby og Gåserød, der vil deltage i lotteriet i den anden by. Når man sammenlægger de to mængder, vil der være nogle som tælles dobbelt; nemlig de personer, der deltager i begge lotterier. For at få regnestykket til at gå op, må man fratrække disse.

Tænk på dette lidt anderledes. Tag et stykke papir og klip to cirkler, der repræsenterer henholdsvis mængderne **A** og **B.** Læg dem på et bord, og lad dem have et overlap.

Overlappet er lig med fælleshændelsen. Overlappet er da også lig med det antal personer, der tælles dobbelt, da disse deltager i lotteriet i såvel Andeby som Gåserød. For at undgå dette, skal de trækkes fra. Det svarer til at man klipper i én af mængderne, og lægger det på bordet således, at der ikke er overlap.

Additionsreglen kan hurtig blive kompliceret! Eksempelvis kan man introducere en tredje by i systemet, og lade der være lotterier i alle byer, hvor borgerne i alle byerne kan deltage i hinandens lotterier. Kaldes de tre byer **A**, **B** og **C** fås *foreningshændelsen* nu som:

$$\begin{aligned} P(A \cup B \cup C) = & P(A) + P(B) + P(C) \\ & - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ & + P(A \cap B \cap C) \end{aligned}$$

Prøv selv at gentage øvelsen med at klippe de tre mængder i papir. Da vil man se, hvorfor der skal stå et plus til sidst! Klipper man i alle mængder vil der være et hul til sidst.

Ovenfor er det blev vist, at der ved *uafhængighed* mellem **A** og **B** gælder, at $P(A \cap B) = 0$. Dette kan anvendes til at opstille *reglen for multiplikation*. Under uafhængighed gælder, at *fælleshændelsen* er lig med:

$$P(A \cap B) = P(A)P(B)$$

Sandsynligheden for fælleshændelsen opnås da ved multiplikation af hændelserne **A** og **B**. Dette er en vigtig regel, når der skal udledes fordelinger, som det vil fremgå af de følgende afsnit.

Ovennævnte er ganske teoretisk. I det følgende vises ved et eksempel, hvordan nogle af termerne kan anvendes til løsning af en opgave.

Eksempel om uddannelsesprogrammer

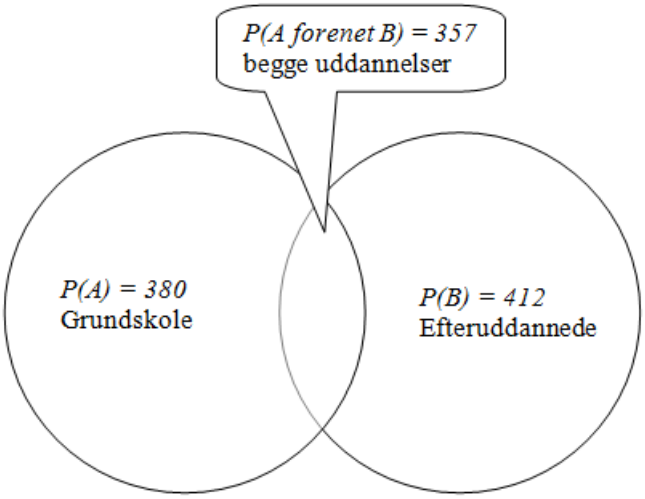
En virksomhed har i alt 550 ansatte. Af disse havde 380 en grundskoleuddannelse, mens 412 havde gennemgået virksomhedens efteruddannelsesprogram. I alt 357 af de ansatte havde både grundskole og efteruddannelse.

Hvis en ansat udvælges tilfældigt, hvad er sandsynligheden for at:

A. Personen har grundskoleuddannelse?
B. Personen har deltaget i efteruddannelsesprogrammet?
C. Personen har enten en grundskoleuddannelse eller en efteruddannelse?

Løsning
Der er to typer af uddannelse nemlig grundskole og efteruddannelse. Lad disse to typer blive betegnet af mængderne **A** og **B.** Endelig er der foreningshændelsen mellem de to grupper.

Lav derfor indledningsvis et *Venn-diagram*, som kan have nedenstående udseende:



Man kan nu beregne sandsynlighederne for at være med i hver gruppe som:

A. *P(A):* Personer med grundskole 380 → P(A) = 380/550 = 0.691
B. *P(B):* Personer med efteruddannelse 412 → P(B) = 412/550 = 0.749

Totalen 550

Det kan jo virke lidt forvirrende, at disse to sandsynligheder summerer til mere end ét, men det er jo netop pointen, da foreningshændelsen resulterer i et overlap svarende til fælleshændelsen.

Det betyder, at der må være en fællesmængde svarende til dem, der har såvel en ene som den anden uddannelse:

$P(A \cap B)$: Personer med begge uddannelser 357

Vi skal nu finde foreningsmængden og anvender "additionsreglen" som før:

C.
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$= \frac{380}{550} + \frac{412}{550} - \frac{357}{550} = 0.691 + 0.749 - 0.649 = 0.791$$

Svarende til 435 af alle ansatte.

2. Trædiagrammer

Et trædiagram kaldes også et *tælletræ*. Her beregnes sandsynligheder ud fra opstilling af informationer om hændelser. Disse vil ofte være *betingede* af hinanden. Eksempelvis *hvad* sker der, *givet* at der initialt er sket noget *andet*. En betinget sandsynlighed udtrykkes med en lodret streg. Eksempelvis:

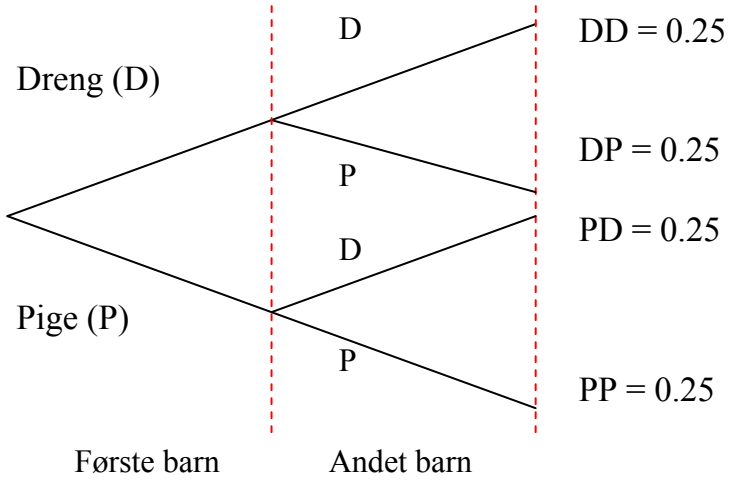
$P(A|B)$: Sandsynligheden for A givet at B allerede er indtruffet

På denne måde fastlægges *udfaldsrummet* for den betingede hændelse. Husk at hvis to hændelser er uafhængige, så kan sandsynlighederne multipliceres med hinanden.

Anvendelsen af trædiagrammet kan bedst illustreres ved et eksempel. Betragt et ægtepar, som gerne vil have to børn. Antag at man ikke kan få tvillinger. Hvad er sandsynligheden for, at de får henholdsvis to piger (P), to drenge (D) eller ét barn af hvert køn?

Biologisk er sandsynligheden for at få en dreng eller en pige lige stor, så der gælder P(P) = P(D) = 0.5, idet sandsynlighederne summerer til 1.

Det samlede udfaldsrum kan fastlægges i et trædiagram, som skitseret nedenfor. Sandsynlighederne fremkommer ved anvendelse af multiplikationsreglen, da det antages at udfaldet af det første og andet barn er uafhængige.



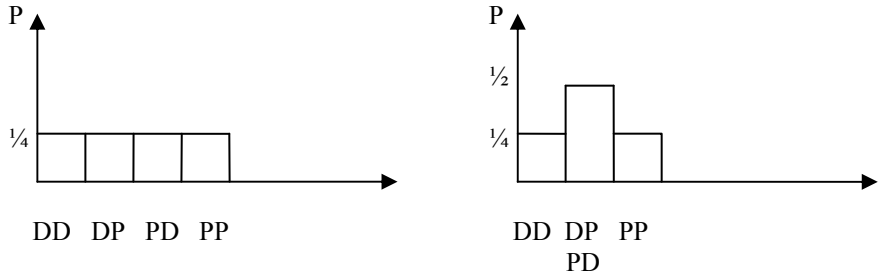
Sandsynligheden for at det første barn bliver en dreng er lig 0.5. Nu er kønnet på det andet barn betinget af kønnet af det første barn. Lad det blive en dreng. Da bliver sandsynligheden for 2 drenge lig med 0.5 gange 0.5 lig med 0.25. De mulige udfald kan nu opstilles som følger:

DD DP PD PP

Der vil således være 25 procent sandsynlighed for to drenge, 25 procent sandsynlighed for to piger, og 50 procent sandsynlighed for en dreng eller pige.

Der gælder da at: $P(DD) + P(DP) + P(PD) + P(PP) = 1.00$

Man kan opstille udfaldsrummet som vist nedenfor. I opstillingen til venstre er udfaldene identiske. Dette benævnes en *uniform* fordeling, mens opstillingen til højre er en symmetrisk fordeling – og en meget grovkornet *normalfordeling.* Dette vendes der tilbage til senere i dette sæt af noter.



Eksempel (eksamen februar 2011, 10 %, 2P)
I et land har 70 procent af de stemmeberettigede vælgere deltaget i parlamentsvalget. 20 procent af de stemmeberettigede, som har deltaget i valget, har stemt på partiet XYZ.

Hvor stor er sandsynligheden for, at en tilfældig udvalgt stemmeberettiget i dette land ganske vist har deltaget i valget, men *ikke* har stemt på partiet XYZ?

Løsning:
Dette er en opgave i sandsynlighedsregning! Vi antager uafhængighed, og må multiplicere sandsynlighederne. Først opskrives informationer, så man ved hvad man arbejder med:

- Valgdeltagelsen var på 70 % af alle stemmeberettigede. Lad os kalde denne for $P(V) = 0.7$

Nu opstilles der betingede sandsynligheder:

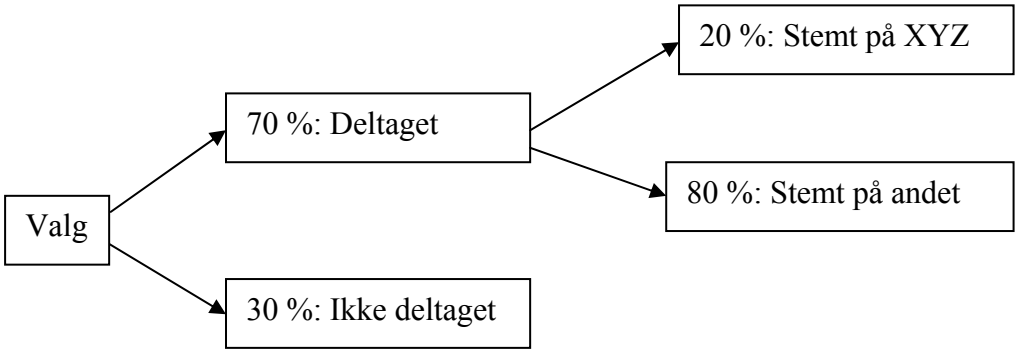
- Givet man har deltaget i valget og stemt på XYZ dvs. $P(XYZ \mid V) = 0.2$

- Har man deltaget i valget og ikke stemt på XYZ fås $P(\text{ikke-XYZ} \mid V) = 1 - 0.2 = 0.8$

(dette er sandsynligheden for at have stemt på et andet parti, men vi mangler, at en tilfældig udvalgt person kan have undladt at stemme, men er stemmeberettiget).

$P(V \text{ og ikke-XYZ} \mid V) = P(V)P(\text{ikke-XYZ} \mid V) = 0.7 \times 0.8 = 0.56$

Ovenstående virker jo rodet, så alternativt kan man anvende et sandsynlighedstræ:



Dvs. at for at deltage, men ikke at have stemt på XYZ er da lig $0,8 \times 0,7 = 0,56$

3. Introduktion til statistiske fordelinger

Verden omkring os er fuld af begivenheder, hvor resultatet ikke er kendt på forhånd for eksempel:

- Hvordan er folketinget eller landdagen sammensat efter næste valg?
- Hvor gammel bliver jeg?
- Får parret i eksemplet ovenfor først en dreng eller en pige?
- Kan fætter Højben vinde lotteriet i både Andeby og Gåserød?

Alle disse forhold kan samles under betegnelsen et *stokastisk forsøg,* der kan defineres som *en aktivitet med to eller flere udfald, hvor udfaldet er usikkert.* Til denne usikkerhed kan der tilordnes en sandsynlighed for at en given hændelse indtræffer.

I eksemplet med parret, som vil have to børn, er der fire muligheder eller stokastiske forsøg, hver med en given sandsynlighed.

En *stokastisk variabel* er en variabel, hvis værdi bestemmes af et stokastisk forsøg, således at hvert udfald karakteriseres ved værdien af den stokastiske variabel.

De fire sandsynligheder, der er skrevet op øverst på side 9 angiver sandsynligheder og udfaldsrum for, at få to børn.

I dette eksempel er det let at opskrive sandsynligheder og udfaldsrum, da antallet af udfald er begrænset. Dette er ikke altid tilfældet. Spørger man sig selv, hvor gammel man bliver, så er der mange muligheder. De fleste statistiske kontorer udregner restlevetiden for en given

aldersgruppe ved anvendelse af sandsynligheder for overlevelse. Men man kan blive kørt over af en damptromle, der kører forbi lige udenfor Munktoft i morgen tidlig, eller man kan gå væk af alderdom i en alder af 105 år. Der er således mange muligheder.

Dette leder til en opdeling af de stokastiske variabler i to klasser:

1. **Diskret stokastisk variabel**, som kun kan antage et endeligt eller tælleligt antal udfald
2. **Kontinuert stokastisk variabel**, som kan antage et uendeligt antal udfald

I de følgende to afsnit vil der blive introduceret to statistiske fordelinger dels binomialfordelingen dels normalfordelingen. Førstnævnte er en diskret fordeling, mens sidstnævnte er en kontinuert fordeling. Nogle har måske stiftet bekendtskab med disse fordelinger i gymnasiet.

Endelig vil der i det sidste afsnit af disse noter blive set på sammenhængen mellem de to fordelinger.

De to fordelinger er de mest anvendte repræsentanter for de to typer af sandsynlighedsfordelinger for stokastiske variabler. Der findes en række andre især diskrete fordelinger som eksempelvis Poissonfordelingen og den multinominale fordeling Disse fordelinger vil ikke blive omfattet af dette kursus.

4. Binomialfordelingen

I denne fordeling ser man på det simplest mulige nemlig en situation med kun to udfald – succes eller fiasko. Det kan eksempelvis være:

- Får parret i eksemplet en først pige eller en dreng?
- Bliver der lavet mål i det næste angreb som Flensburg Handewitt har i håndboldkampen?
- Går bilen i stykker, når man har inviteret kærresten i byen?

En Bernoulli stokastisk variabel med ét forsøg
Situationen med to udfald – succes eller fiasko blev først beskrevet af den svejtsiske matematiker Jakob Bernoulli (1654−1705). Han betragtede en stokastisk variabel X med to udfald nemlig "succes" kaldet 1, som tilordnes sandsynligheden p og "fiasko", kaldet 0, som tilordnes værdien $(1-p)$. Bemærk, at reglen fra afsnit 2 overholdes, der siger, at summen af sandsynlighederne er lig med 1.

Fordelingen af et enkelt af Bernoulli's eksperimenter er givet som:

X	$P(X)$
1 (succes)	p
0 (fiasko)	$1-p$

Det er samme situation, som var gældende i første runde for parret, som gerne vil have et barn – dreng eller pige! Man kan beregne forventningen kaldet E, og variansen kaldet V på følgende måde:

$$\begin{aligned} E(X) &= 1\times p + 0\times(1-p) &&= p \\ E(X^2) &= 1^2\times p + 0^2\times(1-p) &&= p \\ V(X) &= E(X^2) - [E(X)]^2 = p - p^2 &&= p(1-p) \end{aligned}$$

Ofte betegnes $(1-p)$ med q. Så bliver variansen lig med $V(X) = pq$.

Beregningen bygger på de formler, der blev udledt om middelværdien og variansen i det første sæt af noter. For middeltallet anvendes en vægtning, mens der for variansen anvendes en variant af formlen fra notesæt 1 side 24 nederst.

En Bernoulli stokastisk variabel med gentagne forsøg
Forsøget ovenfor er voldsomt forsimplet. Parret ønskede sig jo 2 børn, og i håndboldkampen ovenfor vil der være mange angreb i løbet af matchen! Derfor er det mere relevant at betragte en situation med flere gentagne forsøg.

Betragt en situation med n forsøg. I eksemplet med parret, der vil have børn er n lig med 2. Det kan også være mere kompliceret, som for eksempel i håndboldkampen. Det vendes der tilbage til!

Se på en mønt, der kastes 5 gange, så $n=5$. Lad krone (H=head) og plat (T=tail) være de 2 udfald. Hvis mønten er fair vil begge udfald være lige sandsynlige. Der vil således gælde, at $p = (1-p) = q = 0.5 = ½$.

Betragt følgende situation:

Hvad er sandsynligheden for at præcis 2 "krone" (H) vil opstå i 5 forsøg?

Først er det nødvendigt at finde ud af, hvor mange måder man kan få 2 H på ud af de 5 forsøg. Med vores notation kan alle kombinationer, hvor med får præcis 2 H opskrives som:

TTTHH	TTHTH	THTTH	HTTTH
TTHHT	THTHT	HTTHT	
THHTT	HTHTT		
HHTTT			

Der er således en total på 10 kombinationer. Næste trin er at finde sandsynligheden for hver specifik kombination. Hver kombination består af 5 hændelser, som er uafhængige af hinanden. Det vil sige, at man kan multiplicere sandsynlighederne. Se på en specifik kombination for eksempel HHTTT, der er summen af hændelserne $ppqqq$. Da forekomsterne af krone (H) og plat (T) ikke påvirker nogle af de 10 kombinationer, vil der gælde, at p^2q^3 er sandsynligheden af en hvilken som helst kombination.

Da der er 10 kombinationer, der har værdien H=2, kan man multiplicere p^2q^3 med 10, hvilket giver $P(H=2) = 10p^2q^3$. Under antagelse af, at mønten er fair gælder der, at $P(H=2) = 10(½)^5 = 0.3125$.

Tilsvarende er det muligt at beregne sandsynligheden for enhver anden hændelse. Bemærk at:

- Sandsynligheden for given række af x succes ud af n forsøg med en sandsynlighed for succes lig med p og en fiasko lig med q er givet som:

$$p^x q^{n-x}$$

- Antallet af måder eller kombinationer, hvor n forsøg resulterer i præcis x successer er lig med:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Dette udtryk kaldes på en lommeregner for **nCr**. På en Texas **TI-84 lommeregner** findes dette ved at taste "MATH". Dernæst "PRB" og så 3:nCr. Udtrykket "!" står for fakultet. Det er den kumulative multiplikative sum af et givet tal. For eksempel er fakultet 4! = 1×2×3×4 = 24.

Betragt et lille eksempel. Der er en gruppe på 10 studenter, som skal vælge et hold til at arrangere julefrokosten på 3 studenter. På hvor mange måder kan dette gøres?

Her anvendes formlen:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} = \binom{10}{3} = \frac{10!}{3!(10-3)!} = \frac{3628800}{6 \times 5040} = 120$$

Der er således 120 kombinationer. På en **TI-84 lommeregner** kan man også finde fakultet. Tallet eksempelvis 10, og så "MATH" → "PRB" → 4:! → ENTER → 3628800. Kombinationerne findes ved først at taste 10, og så "MATH" → "PRB" → 3: nCr → 3 → ENTER → 120.

Binominalfordelingen
Med anvendelse af den notation, som er anvendt, kan Binomialfordelingen opskrives som:

$$P(x) = \binom{n}{x} p^x q^{n-x} \qquad \text{hvor } \binom{n}{x} = \frac{n!}{x!(n-x)!}$$

leddet $\binom{n}{x}$ betegnes antallet af kombinationer (nCr). p er sandsynligheden for succes ved et givet forsøg, mens $q=(1-p)$, er fiaskoen. Endelig er n antallet af forsøg, og x er den stokastiske variabel, som angiver *antallet* af successer.

Middelværdien, variansen og standardafvigelsen for Binomialfordelingen er som givet under Bernoulli eksperimentet, men nu multipliceres med antallet af forsøg n. Således fås:

$$\mu = E(X) = np$$
$$\sigma^2 = V(X) = npq$$
$$\sigma = SD(X) = \sqrt{npq}$$

For Bernoulli eller Binominalfordelingen gælder følgende karakteristika:

- Eksperimentet består n identiske forsøg

- Hvert forsøg har to udfald: *p* (succes) og *q* (fiasko)
- Sandsynlighederne er konstante fra forsøg til forsøg
- Forsøgene er uafhængige

Hvis man trækker en lille stikprøve på *n* elementer fra en stor population vil forudsætningerne være gældende.

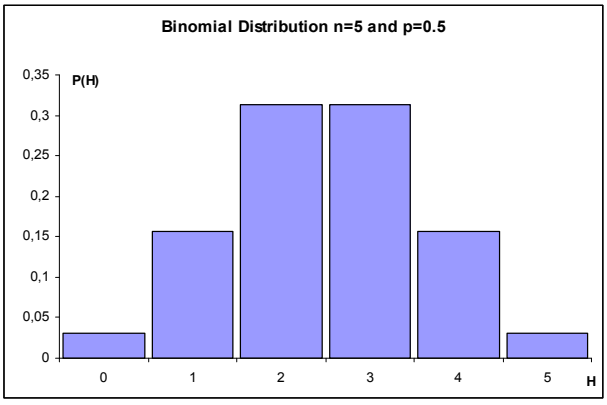
I eksemplet med *n=5* kan sandsynlighederne for den stokastiske variable beregnes til at være:

P(x)	Kombinationer (nCr)	Sandsynligheder	P(H)	Kumulativ P(H)
0	5!/[0!×(5−0)!] = 1	$(½)^0(½)^5 = 0.031$	0.031	0.031
1	5!/[1!×(5−1)!] = 5	$(½)^1(½)^4 = 0.031$	0.156	0.187
2	5!/[2!×(5−2)!] = 10	$(½)^2(½)^3 = 0.031$	0.313	0.500
3	5!/[3!×(5−3)!] = 10	$(½)^3(½)^2 = 0.031$	0.313	0.813
4	5!/[4!×(5−4)!] = 5	$(½)^4(½)^1 = 0.031$	0.156	0.969
5	5!/[5!×(5−5)!] = 1	$(½)^5(½)^0 = 0.031$	0.031	1.000

Det fremgår af tabellen, at fordelingen af symmetrisk jævnfør også illustrationen nedenfor. Man kan beregne middelværdien, variansen og standardafvigelsen som:

$$\mu = E(X) = np = 5 \times ½ = 2.5$$
$$\sigma^2 = npq = 5(0.5)^2 = 1.25$$
$$\sigma = \sqrt{npq} = \sqrt{5(0.5)^2} = 1.12$$

Hvis *p=0.5* er fordelingen symmetrisk. Hvis *p>0.5* er fordelingen venstreskæv. Hvis *p<0.5* er fordelingen højreskæv.



Ovenstående formel kan beregnes på en lommeregner. På en **TI-84 lommeregner** findes sandsynligheden fra eksemplet på følgende måde. Anvend tasten "2nd" og dernæst "DISTR". Under A: Binompdf*(n,p,x)* → "ENTER". Nu indsættes tallene. Det vil sige: A: Binompdf*(5,0.5,2) = 0.3125*. Som forventet.

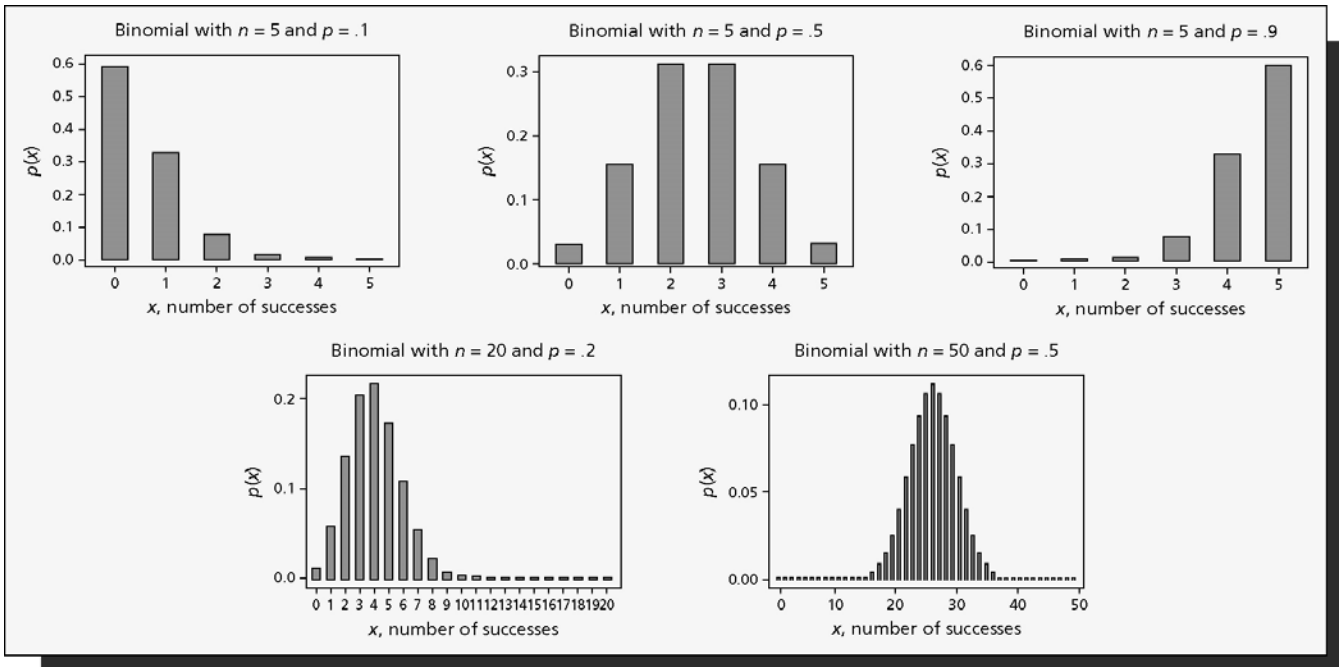
Alternativt anvendes et statistisk tabelværk. Til dette kursus anvendes **Statistics Tables,** som findes i **Blackboard.** Hent pdf-filen med tabellerne og udskriv denne! Tabelværket består af en række tabeller over fordelinger. I statistik I anvendes to af tabellerne, mens de øvrige tabeller anvendes i Statistik II. Alle tabellerne er beregnet med anvendelse af funktioner defineret i Excel.

Den første tabel i **Statistics Tables** omhandler Binomialfordelingen. Midt på side 2 i tabelværket findes nedenstående tabel, hvor *n=5*. I venstre spalte findes *x* og i tabelhovedet findes *p* for *p* op til 0.5. Der læses fra venstre mod højre. Hvis $p > 0.5$ anvendes tabellen omvendt. Så findes *x* i spalten til højre og der læses nedefra og mod venstre.

I tabellen kan sandsynligheden fra eksemplet ovenfor findes som markeret med grøn signatur. Først finder man $n = 5$. Dernæst findes $x=2$ i forspalten. Endelig findes værdien for $p=0.5$ i tabelhovedet. Til sidst aflæses sandsynligheden til at være 0.3125.

n = 5					***P***						
X ↓	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**	
0	0,7738	0,5905	0,4437	0,3277	0,2373	0,1681	0,1160	0,0778	0,0503	0,0313	5
1	0,2036	0,3281	0,3915	0,4096	0,3955	0,3602	0,3124	0,2592	0,2059	0,1563	4
2	0,0214	0,0729	0,1382	0,2048	0,2637	0,3087	0,3364	0,3456	0,3369	0,3125	3
3	0,0011	0,0081	0,0244	0,0512	0,0879	0,1323	0,1811	0,2304	0,2757	0,3125	2
4	0,0000	0,0005	0,0022	0,0064	0,0146	0,0284	0,0488	0,0768	0,1128	0,1563	1
5	0,0000	0,0000	0,0001	0,0003	0,0010	0,0024	0,0053	0,0102	0,0185	0,0313	0
	0,95	**0,90**	**0,85**	**0,80**	**0,75**	**0,70**	**0,65**	**0,60**	**0,55**	**0,50**	**X ↑**

Betragt afslutningsvis nogle grafiske præsentationer af nogle Binomialfordelinger:



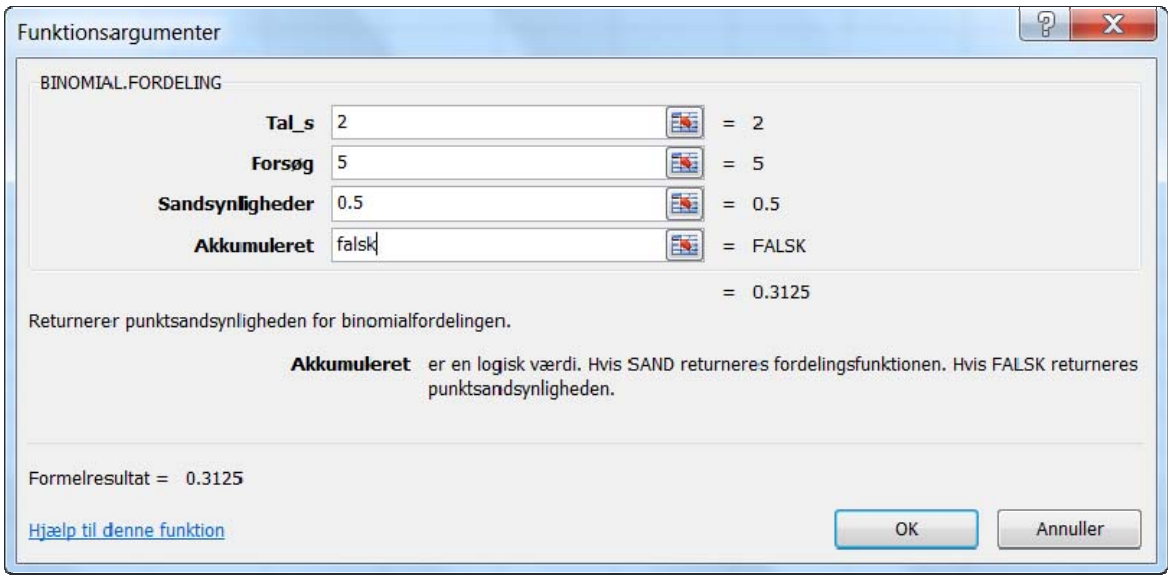
Af illustrationerne kan man se at:

- At Binomialfordelingen bliver mere *symmetrisk* når n øges og når $p \to 0.5$.
  (Nedre højre illustration)
- At Binomialfordelingen bliver *højreskæv* når $p < 0.5$.
  (Øverste venstre illustration)
- At Binominalfordelingen bliver *venstreskæv* når $p > 0.5$.
  (Øverste højre illustration)

Når n øges vil Binomialfordelingen tilnærmes Normalfordelingen. Dette vil blive behandlet i afsnit 6 i disse noter.

Binomialfordelingen kan også findes i Excel ved sekvensen:

Formler | Indsæt funktion | statistisk | binomialfordeling



På denne måde kan man finde sandsynligheden i eksemplet ovenfor. Man kan finde såvel den diskrete som den kumulative sandsynlighed. Man har følgende muligheder:

- Den diskrete sandsynlighed $P(X=4)$ ("Falsk" i Excel)
- Den kumulative sandsynlighed $P(X \leq 4)$ ("Sand" i Excel)

5. Normalfordelingen

Normalfordelingen er den vigtigste af de kontinuerte fordelinger. Rigtig mange økonomiske variabler kan analyseres ved anvendelse af Normalfordelingen, der er en meget robust fordeling, som det også vil fremgå af dette, og det næste sæt af noter, hvor *den centrale grænseværdisætning* udledes.

Normalfordelingen kaldes også for *Gauss Fordelingen* efter den tyske matematiker Carl Friedrich Gauss[1] (1777−1855). Gauss viste, at den matematiske funktion, som er gengivet nedenfor, var den bedst mulige til at tilnærme de tilfældige fejl, der opnås i regressionsanalyse. Der vendes tilbage til dette emne i notesæt 5 i Statistik II.

Gauss er en helt central figur i udviklingen af den moderne matematik og statistik. Han var i Nobelprisklassen og har optrådt både på de gamle D-mark sedler som på frimærker fra det gamle DDR!

Gauss anvendte i 1802 regression til meget præcist at estimere og forudsige banen for den nyligt opdagede asteroide *Ceres´*, der er i omløb mellem Jorden og Mars.

Tæthedsfunktionen *f(x)* for en kontinuert stokastisk variabel med middelværdien μ og standardafvigelsen σ, er givet ved følgende komplicerede udtryk:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \qquad \text{for } -\infty < x < +\infty$$

Hvor e er den naturlige logaritme. Kombinationen af denne samt tallet π sikrer den symmetriske form, der er gengivet i illustrationen i det følgende.

[1] Nok var Gauss en banebrydende og genial matematikker, men han var også en besværlig natur, vanskelig at omgås og forlod meget sjældent Göttingen. Gauss levede på samme tid, som en anden banebrydende tysk videnskabsmand, nemlig Alexander von Humboldt, som rejste over det meste af verden og indsamlede planter, studerede folkeslag mm. Den tyske forfatter Daniel Kehlmann har skrevet en meget underholdende bog om de herrer med titlen "Opmålingen af verden", (Forlaget Per Kofoed, 2009 (Tysk titlel: "Die Vermessung der Welt" (Rowohlt Verlag, 2005)), hvor han sætter de to personligheder overfor hinanden i en periode præget af politiske forandringer. Bogen har solgt i flere millioner eksemplarer og er oversat til de fleste europæiske sprog. Anbefales, hvis man kører sur i noterne til statistik og mikroøkonomi! Filmatiseret med premiere oktober 2012, som en adventurefilm i 3D instrueret af Detlev Buck.

I den teoretiske form har normalfordelingen middelværdien lig med nul, og variansen lig med ét. Da kvadratroden af ét er lig med ét, er varians og standardafvigelsen lig med hinanden. Dette skrives ofte som:

$$Z \approx N(0,1^2)$$

For en stokastisk normalfordelt variabel *x* skrives ofte:

$$X \approx N(\mu_x, \sigma_x) \qquad \text{eller} \qquad X \approx N(\mu_x, \sigma_x^2)$$

Tæthedsfunktionen for *Z* i formlen på den forrige side er for den teoretiske idealsituation. For at komme fra denne til et observeret datasæt *x* anvendes følgende vigtige udtryk for transformationen mellem *x* og *Z:*

$$Z = \left(\frac{X - \mu_x}{\sigma_x} \right)$$

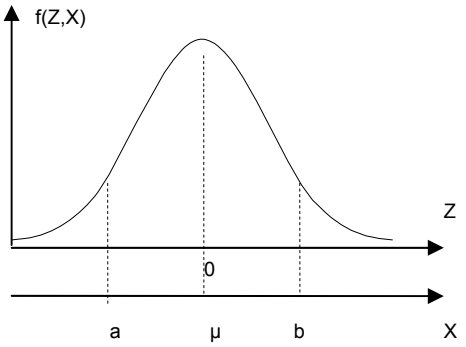
Formlen transformerer den stokastiske variabel *x* med middelværdien μ_x og standardafvigelsen σ_x til en normalfordelt variabel.

Denne formel gør det muligt for givne værdier *a* og *b*, at beregne til tilhørende sandsynligheder svarende til arealer under kurven for standardnormalfordelingen. Følgende formler meget anvendelige kan opstilles for beregning af sandsynligheder:

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

$$P(X > b) = P\left(Z > \frac{b - \mu}{\sigma} \right)$$

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



Transformationsformlen kan også opskrives *inverst.* Her findes for en given sandsynlighed en specifik værdi af *x*. Den inverse transformation ser ud som følger:

$$X = \mu_x + Z\sigma_x$$

Ganske som tidligere kan sandsynligheder enten findes ved anvendelse af lommeregneren eller ved brug af tabellen i **Statistics Tables** på side 9.

Eksempel: Prisen på benzin
I Danmark har der i mange år været priskrig på benzin. Mens listeprisen som regel er højere end i Tyskland, så er tilbudsprisen lavere end i Tyskland. Benzin forhandlet i eksempelvis Kruså, Sønderborg eller Aabenraa er ofte på tilbud især i morgentimerne og specielt om mandagen. Som regel hæves prisen kl. 10:00, for så at falde igen senere på dagen. Skal man eksempelvis tanke 40 liter kan man på tilbud finde prisen op til 1.50 DKK lavere, så man kan spare 60 DKK på en optankning – det er værd at køre efter!

I foråret 2006 indsamlede forfatteren af disse noter priser på benzin for 95 oktan. Ved 25 besøg på benzinstationer fandtes en middelværdi på 9.95 DKK per liter med en standardafvigelse på 0.30 DKK per liter. Antag nu, at prisen på benzin er Normalfordelt med disse værdier.

Besvar nu følgende:

A. Prisen på benzin er en tilfældig dag lig med 10.40 DKK per liter. Hvad er sandsynligheden for, at prisen vil være højere ved bilistens næste besøg på en benzinstation?
B. Engang var bilisten heldig og kunne købe benzin til pris på 9.65 DKK per liter. Hvad er sandsynligheden for, at prisen vil blive lavere ved bilistens næste besøg på en benzinstation?
C. Hvad er sandsynligheden for, at bilisten vil finde en pris på benzin mellem 9.80 DKK og 10.25 DKK ved det næste besøg på en benzinstation?
D. Antag nu, at bilisten forventer en lavere pris 25 % af de gange, som denne besøger en benzinstation. Hvad er minimumsprisen bilisten vil acceptere ved den givne middelværdi og standardafvigelse?

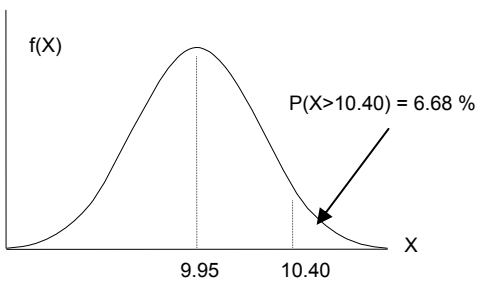
Løsning
Man betragter en normalfordelt stokastisk variabel med *N(9.95;0.30)*. Størrelsen af datasættet er lig med *n=25* (denne information vil blive anvendt i de følgende sæt af noter).

A)
Prisen på benzin er en tilfældig dag lig med 10.40 DKK per liter. Hvad er sandsynligheden for, at prisen vil være højere ved bilistens næste besøg på en benzinstation?

Til at løse denne problemstilling indsættes i den midterste formel ovenfor:

$$P(X > 10.40) = P(\frac{X-\mu}{\sigma} > \frac{10.40-9.95}{0.30}) = P(Z > 1.5) = 1 - 0.9332 = 0.0668\ (= 6.68\ \%)$$

Illustration:



Den lille illustration ovenfor skal man prøve at lave helst hver gang. Det letter overblikket, og gør det lettere at løse en given problemstilling.

Hvordan fremkommer sandsynligheden? Beregningen af *Z* skulle ikke volde kvaler! Men hvordan finder man den sandsynlighed, som passer med *Z*? Denne kan enten findes på lommeregneren eller i **Statistics Tables.**

På side 9 i **Statistics Tables** findes en tabulering af den kumulerede normalfordeling. Vi skal finde den sandsynlighed, der passer til Z = 1.5. I tabellens forspalte finder man den værdi, der passer til 1.5. Så går man ind i tabellen under "0.00", da anden decimal er lig med 0, og finder den tilhørende sandsynlighed, som er 0.9332. Nu skulle man finde sandsynligheden, når prisen er *højere* end 10.40 DKK. Da det vides at arealet under hele normalfordelingskurven er lig med 1, fratækkes den netop fundne sandsynlighed, og man får svaret som 0.0668.

Z	0,00	0,01	0,02	0,03	0,04	…	0,09
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	…	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	…	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	…	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	…	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	…	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	…	0,9706

Betragt et andet eksempel på opslag. Hvad nu hvis Z = 1.73? Her finder man først Z=1.7 i forspalten. Nu skal man finde den anden decimal, som er lig med 3. Man går vandret ind i tabellen og finder værdien "0.03". Dernæst går man ned til Z=1.7. Her findes sandsynligheden til at være lig med 0.9582.

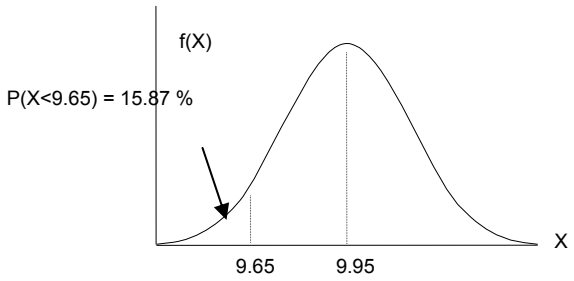
På en **TI-84 lommeregner** kan også beregne sandsynligheden. Anvend tasten "2nd" og dernæst "DISTR". Under 2: Normalcdf*(low,high,μ,σ)* → "ENTER". Her er *low* den lavere grænse, og *high* er den øvre grænse.

I spørgsmålet ovenfor vil det se ud som følger: 2: Normalcdf*(0,10.40,9.95,0.30)* = 1−0.9302 = 0.068. Det vil sige, at man beregnes sandsynligheden op til 10.40 DKK. Den er lig med 0.9302. Dette tal fratrækkes 1, hvorved det rigtige svar fremkommer.

B)
Engang var bilisten heldig og kunne købe benzin til pris på 9.65 DKK per liter. Hvad er sandsynligheden for, at prisen vil blive lavere ved bilistens næste besøg på en benzinstation?

$$P(X < 9.65) = P(\frac{X-\mu}{\sigma} < \frac{9.65-9.95}{0.30}) = P(Z < -1) = 0.1587\ (=15.87\,\%)$$

Illustration:

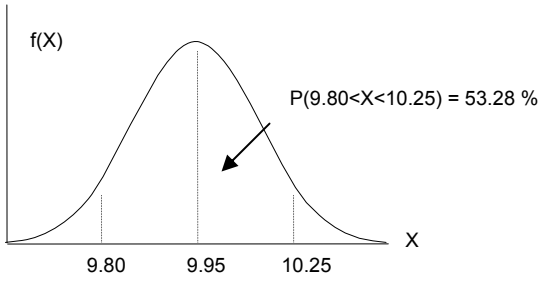


Tabellen i **Statistics Tables** anvendes som skitseret ovenfor. På **TI-84 lommeregneren** fås følgende: "2nd" og dernæst "DISTR" → 2: Normalcdf*(0,10.40,9.95,0.30)* = 0.1587.

C)
Hvad er sandsynligheden for, at bilisten vil finde en pris på benzin mellem 9.80 DKK og 10.25 DKK ved det næste besøg på en benzinstation

$$P(9.80 < X < 10.25) = P(\frac{9.80-9.95}{0.30} < \frac{X-\mu}{\sigma} < \frac{10.25-9.95}{0.30}) = P(-0.5 < Z < 1) = 0.8413 - 0.3085 =$$
$$0.5328\,(=53.28\,\%)$$

Illustration:



Tabellen i **Statistics Tables** anvendes som skitseret ovenfor. På **TI-84 lommeregneren** fås følgende: "2nd" og dernæst "DISTR" → 2: Normalcdf*(9.80,10.25,9.95,0.30)* = 0.5328.

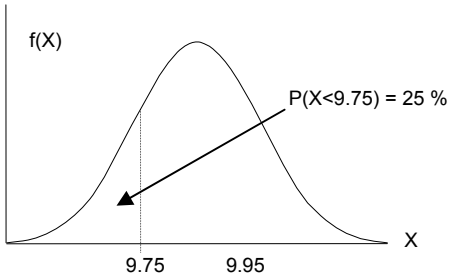
D)
Antag nu, at bilisten forventer en lavere pris 25 % af de gange, som denne besøger en benzinstation. Hvad er minimumsprisen bilisten vil acceptere ved den givne middelværdi og standardafvigelse?

I dette tilfælde løses der for *a,* idet sandsynligheden allerede er givet til 25 %. Man får da:

$$P(X < a) = P(\frac{a - 9.95}{0.30}) = 0.25 \Leftrightarrow a = 9.95 + (-0.67)0.30 \Leftrightarrow a = 9.75$$

Minimumsprisen blisten vil acceptere er lig med 9.75 DKK

Illustration



Spørgsmålet er nu, hvordan dettes gøres, og hvordan man finder værdien −0.67. Dette er værdien for *Z,* når sandsynligheden er 25 % svarende til 0.25.

Man anvender tabellen for Normalfordelingen på side 9 i **Statistics Tables** ”inverst”, som gengivet nedenfor. Først finder man ”inde i tabellen” den nærmeste værdi til 0.25. Det er i dette tilfælde værdien 0.2514. Læser man vandret finder man at Z = −0.6. Dernæst læses lodret. Her findes værdien 0.07, som lægges til de 0.6, så totalen bliver Z = −0.67.

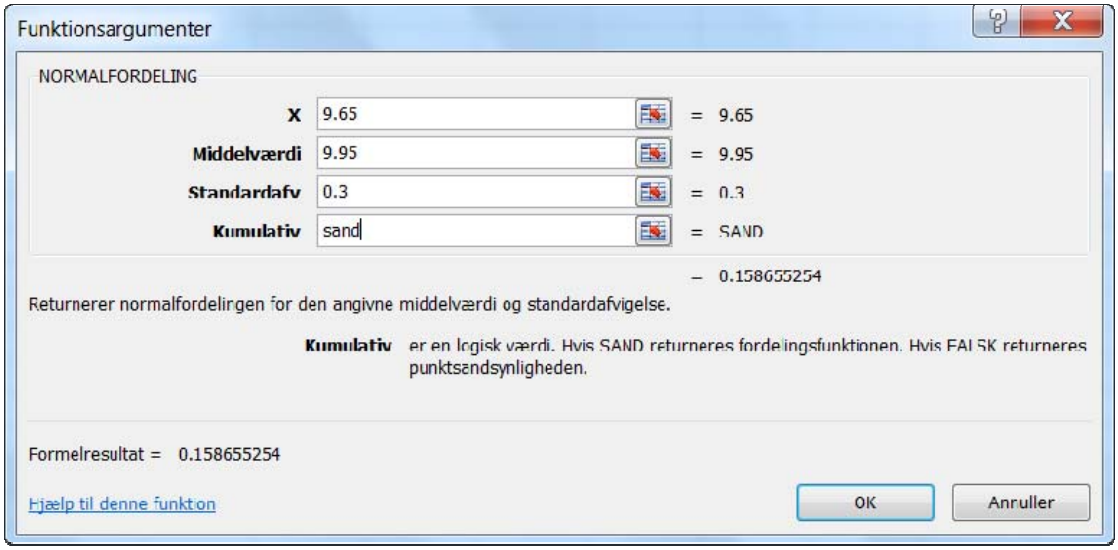
Z	0,00	…	0,05	0,06	0,07	0,08	0,09
-0,7	0,2420	…	0,2266	0,2236	0,2206	0,2177	0,2148
-0,6	0,2743		0,2578	0,2546	0,2514	0,2483	0,2451
-0,5	0,3085	…	0,2912	0,2877	0,2843	0,2810	0,2776
-0,4	0,3446	…	0,3264	0,3228	0,3192	0,3156	0,3121

Som tidligere kan man finde sandsynligheden ved anvendelse af **TI-84 lommeregneren.** Brug “2nd” og dernæst “DISTR”. Under 3: InvNorm*(area,μ,σ)*. Hvor *area* betegner sandsynligheden. I eksemplet fås 3: InvNorm*(0.25,9.95,0.30)* = 9.7476 ≈ 9.75. Som forventet.

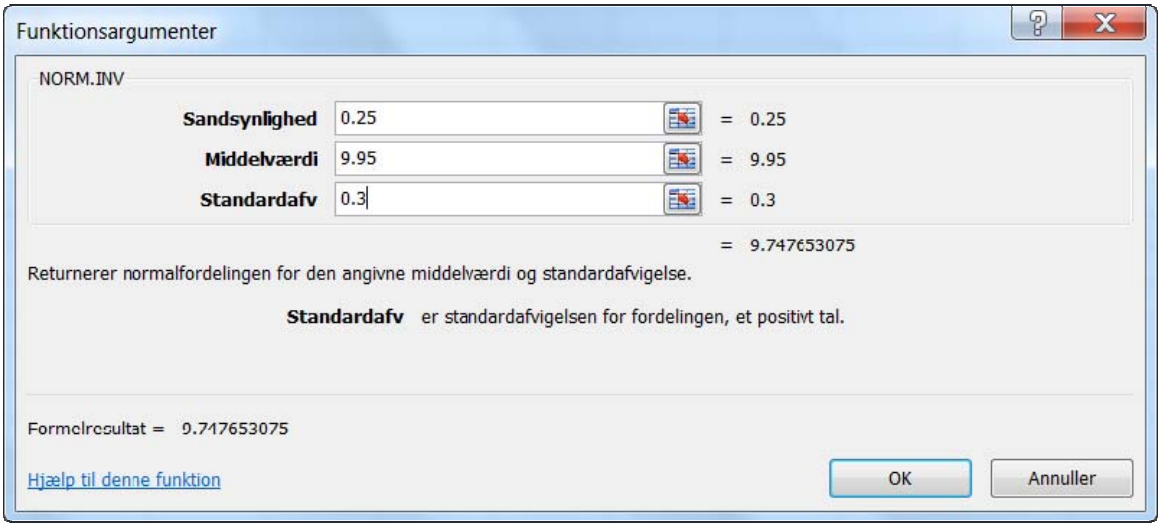
Som det var tilfældet med Binomialfordelingen kan man med hensyn til Normalfordelingen også anvende **Excel** til at beregne sandsynlighederne ved sekvensen:

Formler | Indsæt funktion | statistisk | Normalfordeling

For spørgsmål *B* i eksemplet kan sandsynligheden findes som:



Tilsvarende kan man finde den inverse sandsynlighed, som anvendes i spørgsmål *D* ved brug af sekvensen **Formler | Indsæt funktion | statistisk | Norm.inv**



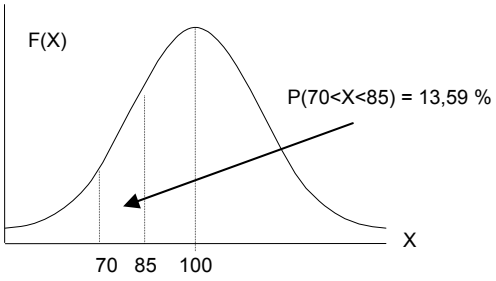
Eksempel (eksamen februar 2011, 15 %, 3P)
Intelligenskvotienten, målt med en bestemt testmetode, er i befolkningen normalt fordelt med en middelværdi på 100 og en standardafvigelse på 15 point. Man taler om en "under gennemsnitlig intelligens" (dog endnu ikke om "evnesvaghed") ved en IQ-værdi på mellem 70 og 85. Hvor stor en procentdel af befolkningen ville denne beskrivelse passe på?

Løsning:
Det antages at materialet er normalfordelt og $\mu = 100$ mens $\sigma = 15$.

Vi skal finde, hvor stor en procentandel af data sættet som har "gennemsnitlig intelligens" mellem 70 0g 85. Dvs. at vi skal finde:

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

Illustration*:*



Ved indsættelse og anvendelse af en tabel for normalfordelingen fås:

$$P(70 < X < 85) = P(\frac{70-100}{15} < \frac{X-\mu}{\sigma} < \frac{85-100}{15}) = P(-2 < Z < -1) = 0.1587 - 0.0228 =$$
$$0.1359\,(=13.59\,\%)$$

Kan findes direkte i en **TI-84 lommeregner** som: "2nd" → "DISTR" → 2: Normalcdf(70,85,100,15) = 0.1359

6. Sammenhængen mellem Normal- og Binominalfordelingen

Vi har tidligere i disse noter beregnet middelværdien og standardafvigelsen for en binomialfordel stokastisk variabel. Hvis n betegner antallet af forsøg, mens p er sandsynligheden for succes og q er sandsynligheden for fiasko, så fandtes det at:

$$E(X) = \mu = np \qquad \text{og} \qquad SD(X) = \sqrt{npq}$$

Denne information kan anvendes i de udtryk, der fandtes ovenfor, for beregning af sandsynligheden for en stokastisk normalfordelt variabel. Ønsker man eksempelvis at beregne sandsynligheden for at have et udfald mellem to værdier benævnt *a* og *b* fås ved indsættelse:

$$P(a < X < b) = P\left(\frac{a - np}{\sqrt{npq}} < Z < \frac{b - np}{\sqrt{npq}} \right)$$

Dette holder for en Binomialfordelt stokastisk variable med $n > 20$ sandsynligheden p i intervallet mellem $0.1 \leq p \leq 0.9$.

Nils Karl Sørensen ©
Statistik I, BA int, Flensborg

Sæt 3: Estimation og konfidensintervaller

af Nils Karl Sørensen

Indhold *side*

1. Stikprøver og estimatorer

I statistik arbejdes der næsten altid med stikprøver, som er udtrukket fra en totalpopulation. I forbindelse med processen omkring udtrækningen af en stikprøve, kan der rejses en række vigtige spørgsmål:

- Hvis totalpopulationen er Normalfordelt, vil stikprøven da også være Normalfordelt?
- Hvad sker der med middelværdien og standardafvigelsen i stikprøven? Vil disse værdier være som i totalpopulationen?
- Hvad nu hvis totalpopulationen ikke er Normalfordelt? Vil stikprøven da være Normalfordelt?
- Vil en stor stikprøve være bedre end en lille stikprøve?
- Hvilke krav skal man stille til en stikprøve, for at den er god?

Hvordan man IKKE tager en stikprøve!

Lad os begynde diskussionen med, at se på et eksempel på, hvordan man ikke tager en stikprøve! Stikprøver anvendes ofte til at forudsige udfaldet af eksempelvis et lokal- eller et nationalt valg. I dag anvendes som regel ”exit pools”, hvor man spørger vælgerne, om deres afgivelse af stemme umiddelbart efter de har stemt. Hvis valgkredsen er repræsentativt sammensat med hensyn til befolkningens stemmeafgivelse, så kan man allerede kort tid efter at valghandlingen er afsluttet, danne sig et rimeligt billede af udfaldet.

Sådan har det imidlertid ikke altid været! Det berømteste eksempel på, hvordan man kan tage grundigt fejl omkring udfaldet af en valghandling er nok den forkerte forudsigelse af udfaldet af det amerikanske præsidentvalg i oktober 1936, hvor den demokratiske kandidat F. D. Roosevelt vandt stort over den republikanske udfordrer A. M. Landon.

Tidsskriftet *Literary Digest* havde ved de to tidligere præsidentvalg korrekt forudsagt udfaldet af præsidentvalget, og op til valget i 1936 lavede bladet igen en undersøgelse, der forudsagde en overbevisende valgsejr til A. M. Landon!

Hvad gik der galt? Undersøgelsen af *Literary Digest* gav A. M. Landon sejren i 32 stater med tilsammen 370 valgmænd mod F. D. Roosevelt's som skulle få 161 valgmænd. Valgets udfald var noget anderledes! F. D. Roosevelt vandt en overbevisende valgsejr med en majoritet på godt 11 millioner stemmer. Det er den største forskel, der nogensinde er registreret! F. D. Roosevelt vandt alle stater med undtagelse af Maine og Vermont. Han fik i alt 532 valgmænd mod 8 til A. M. Landon. *Literary Digest* mistede så meget troværdighed, at ingen ville købe tidsskriftet, og virksomheden gik i betalingsstandsning.

Literary Digest prøvede at forudsige valget med udgangspunkt i en stikprøve på 10 millioner vælgere. Man anvendte lister fordelt på stater med abonnenter af bladet, lister af telefonabonnenter og registreringer af automobiler. På det tidspunkt lå ledigheden i USA på omkring 25 procent, og der var ikke nogen form for kompensation. Indkomsterne var således dels lave og dels meget uens fordelt. Ved at vælge telefonlister, registreringer af automobiler og vælgere, der havde råd til et bladabonnement, fik man fat i vælgere med en indkomst. Disse var ikke interesserede i F. D. Roosevelt's økonomiske "New Deal"-program, der ville skabe beskæftigelse ved at stimulere byggeri af infrastruktur som veje, broer og kraftanlæg, og have skatten for de velhavende. Stikprøven havde derfor en skævhed – eller bias – mod A. M. Landon's vælgere. Yderligere var der kun 2.3 millioner, som svarede på undersøgelse. Frafaldet var derfor alt for stort, og dets sammensætning blev ikke undersøgt.

På samme tid forudsagde et på det tidspunkt helt ukendt analytiker ved navn *Gallup* valget korrekt på basis af en stikprøve på kun 2,300 respondenter.

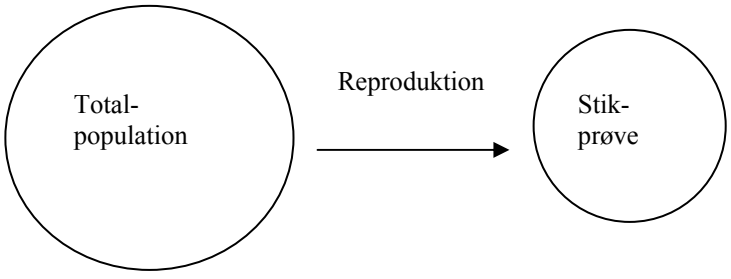
Det amerikanske præsidentvalg har senere hen også været udsat for fejlagtige forudsigelser. På billedet på næste side ser man en situation fra præsidentvalget i 1947, hvor vinderen H. D. Truman holder første udgaven af Chicago Daily Tribune, der fejlagtigt udråber ham som taber af valget til T. E. Dewey.

Estimatorer og deres karakteristika

Hvad er så moralen af dette? En stikprøve skal være den *bedst mulige* og ikke *den størst mulige.* Formålet med at tage en stikprøve er så præcist som mulig at forudsige valgets udfald (position) og usikkerheden på udfaldet typisk repræsenteret ved standardafvigelsen. I tilfældet ovenfor er det stemmeandelen til vinderen og usikkerheden på denne andel.

Efter valghandlingen og stemmerne er optalt, har man al information. Man kan da sammenligne totalpopulationen i form af de afgivne stemmer med den forudsagte prognose. Værdierne fra totalpopulationen benævnes *populations parametrene,* mens værdierne i stikprøven benævnes *stikprøve estimatorerne.*

Skematisk kan estimationsprocessen vises som følger:



Definer på denne baggrund:

- En *estimator* er stikprøvens indikator for en *populations parameter.* Det vil sige for totalpopulationen. Man kan have enten *punkt estimatorer* eller *interval estimatorer*. Sidstnævnte anvendes ofte indenfor marketing, men vil ikke blive omfattet af nærværende sæt af noter

I de følgende sæt af noter anvendes følgende notation omkring parametre og estimatorer:

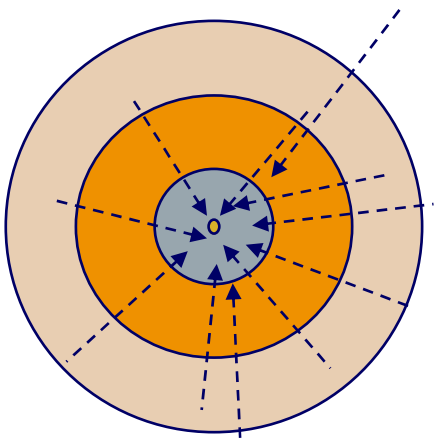
	Estimator (stikprøven)		***Population parameter (totalen)***
Mideelværdien:	$\overline{X} \rightarrow$	estimerer	μ
Variansen:	$s^2 \rightarrow$	estimerer	σ^2
Populationsandelen:	$\hat{p} \rightarrow$	estimerer	p

Tidligere i notesættet om *deskriptiv statistik* defineredes middelværdien og variansen. Nu kan *populationsandelen* defineres som:

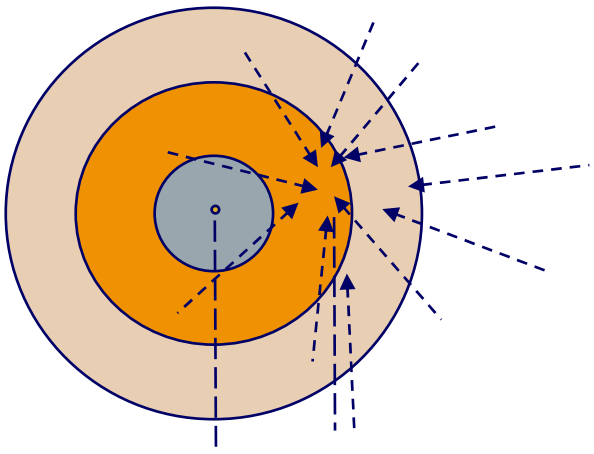
$$\hat{p} = \frac{x}{n}$$

Hvor x er antallet af elementer i stikprøven som tilhører en bestemt kategori eksempelvis antallet af personer, der har stemt "ja" eller personer med præference for et givet produkt. Endelig betegner n stikprøvens størrelse.

Karakteristika for estimatorer



En **unbiased** (ikke skæv) estimator har middelværdi som totalen



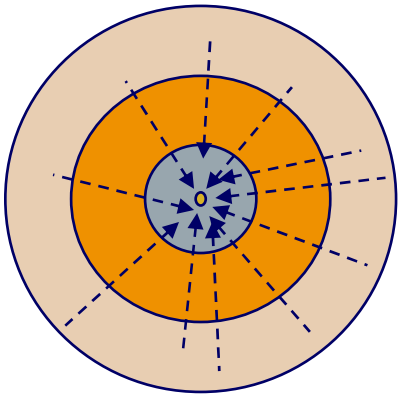
Bias

En **biased (skæv)** estimator har middelværdi forskubbet i forhold til totalen.

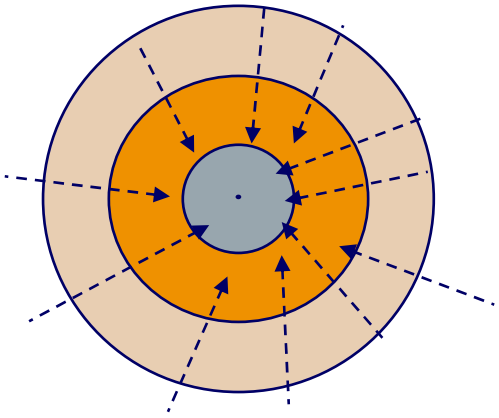
En god estimator skal kunne opfylde følgende karakteristika:

1. En estimator siges at være ikke skæv ***(unbiased),*** hvis den forventede værdi er lig med den totalpopulations parameter, som estimeres. Dette er illustreret på den foregående side. Et eksempel på en skæv estimator er eksemplet med det amerikanske præsidentvalg i 1936.

2. Es estimator siges, at være ***efficient,*** hvis variansen (eller standardafvigelsen) er relativt lille.

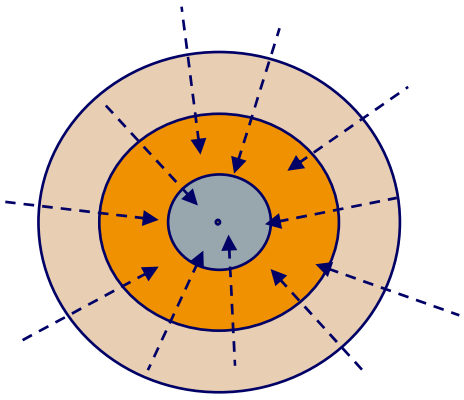


En **efficient** estimator har en lille variation omkring den middelværdi, der beregnes i stikprøven

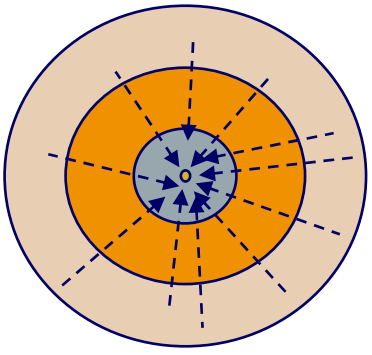


En **ikke-efficient** estimator har en stor variation omkring den middelværdi, der beregnes i stikprøven

3. En estimator siges at være ***konsistent,*** hvis sandsynligheden for i stikprøven at være tæt på middelværdien i totalpopulationen øges med størrelsen af stikprøven



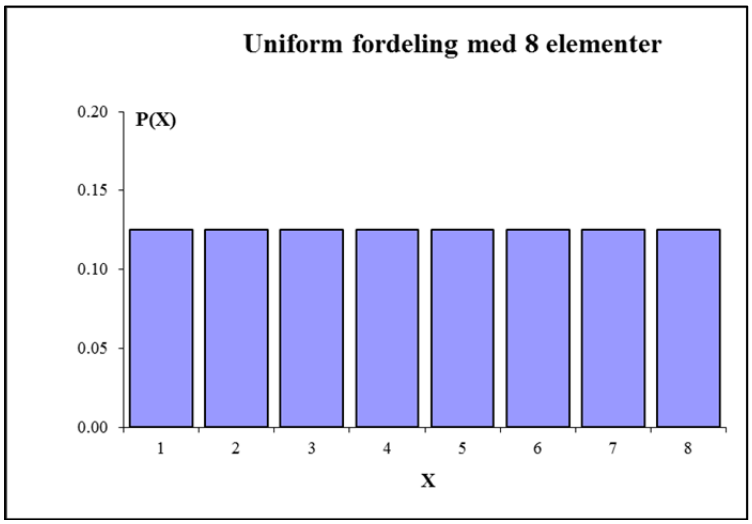
n = 10



n = 100

Den centrale grænseværdisætning
Hvad sker der med estimatorer af middelværdien og variansen, når der tages en stikprøve? Svaret er indeholdt i den centrale grænseværdisætning! Kort sagt, så siger den, at en stikprøve altid vil være Normalfordelt uanset, hvilken fordeling totalpopulationen følger. Yderligere, så vil middelværdierne i stikprøven og totalpopulationen være identiske, mens variansen vil være mindre i stikprøven.

For at vise dette betragtes et eksempel, hvor man tager stikprøver fra en totalpopulation. Dernæst sammenholdes middelværdi og varians for såvel totalpopulationen som for gennemsnittet af alle stikprøverne.



Figuren ovenfor angiver totalpopulationen. Det er et datasæt med 8 udfald, der alle er lige sandsynlige. Dette kaldes for en *uniform fordeling*. Udfaldene går fra 1 til 8, og da de alle er lige sandsynlige, er udfaldssandsynligheden lig med 1/8 = 0.125. Nu beregnes middelværdien μ, variansen σ^2 samt standardafvigelsen σ. Dette gøres i tabellen nedenfor:

X	P(X)	XP(X)	(X−μ)	$(X-\mu)^2$	$P(X)(X-\mu)^2$
1	0.125	0.125	−3.5	12.25	1.53125
2	0.125	0.250	−2.5	6.25	0.78125
3	0.125	0.375	−1.5	2.25	0.28125
4	0.125	0.500	−0.5	0.25	0.03125
5	0.125	0.625	0.5	0.25	0.03125
6	0.125	0.750	1.5	2.25	0.28125
7	0.125	0.875	2.5	6.25	0.78125
8	0.125	1.000	3.5	12.25	1.53125
Sum	1.000	4.500			5.25000

Man får da følgende værdier:

$$E(X) = \mu = 4.5 \qquad V(X) = \sigma^2 = 5.25 \qquad SD(X) = \sigma = 2.2913$$

Af hensyn til eksemplet er værdierne beregnet lidt anderledes end normalt, så tabellen kræver lidt mere forklaring. Alle udfald har samme sandsynlighed. Middelværdien kan da beregnes ved at gange sandsynligheden med værdien af variablen. Det vil side den første kolonne gange med den anden kolonne. Værdien heraf findes i den tredje kolonne. Summeres denne fås middelværdien.

Den fundne middelværdi anvendes til at beregne variansen. Fra noten om *deskriptiv statistik* huskes, at variansen er lig med summen af de kvadrerede afvigelser fra de enkelte observationer til middelværdien. Den netop fundne middelværdi fratrækkes for hver af de 8 observationer og kvadreres. Dette sker i fjerde og femte kolonne i tabellen. Endelig ganges den fundne værdi med sandsynligheden i den sjette kolonne, og der summeres.

For totalpopulationen gælder således, at middelværdien er lig med 4.5, og variansen er lig med 5.25. Endelig er standardafvigelsen lig med kvadratroden af denne, det vil sige 2.2919.

Nu tager vi stikrøver fra denne totalpopulation og sammenligner. Det, som man skal finde, er den *samlede fordeling af alle stikprøver for* $\overline{X}$. For at gøre det så enkelt som muligt, tages der stikprøver af den mindst mulige størrelse. Det vil side, at **n=2.** Først tages der ét element, og dernæst tages det andet element. Begge elementer tages af hele totalpopulationen. Hvor mange kan dette gøres på? Der er 8 elementer i totalpopulationen. Da der er uafhængighed mellem første og anden runde udtagning, kan der tages i alt 8×8 = 64 forskellige stikprøver, der alle er lige sandsynlige.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	(1:1)	(1:2)	(1:3)	(1:4)	(1:5)	(1:6)	(1:7)	(1:8)
2	(2:1)	(2:2)	(2:3)	(2:4)	(2:5)	(2:6)	(2:7)	(2:8)
3	(3:1)	(3:2)	(3:3)	(3:4)	(3:5)	(3:6)	(3:7)	(3:8)
4	(4:1)	(4:2)	(4:3)	(4:4)	(4:5)	(4:6)	(4:7)	(4:8)
5	(5:1)	(5:2)	(5:3)	(5:4)	(5:5)	(5:6)	(5:7)	(5:8)
6	(6:1)	(6:2)	(6:3)	(6:4)	(6:5)	(6:6)	(6:7)	(6:8)
7	(7:1)	(7:2)	(7:3)	(7:4)	(7:5)	(7:6)	(7:7)	(7:8)
8	(8:1)	(8:2)	(8:3)	(8:4)	(8:5)	(8:6)	(8:7)	(8:8)

Alle de 64 stikprøver er gengivet i tabellen. I den første stikprøve trak man tallet 1 i første runde og tallet 1 i anden runde. Derfor står der (1:1). Læser man vandret mod højre, så finder man den næste stikprøve, hvor man i første runde trak tallet 1 og i anden runde tallet 1. Derfor står der (1:2). Og så fremdeles.

Da det ønskes at finde middelværdien for alle de forskellige muligheder at trække stikprøven på, skal man nu beregne middelværdien for hver stikprøve af 2 elementer. Dette gøres i den næste tabel. Betragt et par eksempler: For stikprøven (1:4) findes gennemsnittet,

som summen af observationerne divideret med antallet af observationer: (1+4)/2 = 2.5. Gennemsnittet af stikprøven (8:4) er da lig (8+4)/2 = 6.0 og så fremdeles. Alle middelværdierne er vist nedenfor.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5
2	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
3	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5
4	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0
5	3.0	3.5	4.0	4.5	5.0	5.5	6.0	6.5
6	3.5	4.0	4.5	5.0	5.5	6.0	6.5	7.0
7	4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.5
8	4.5	5.0	5.5	6.0	6.5	7.0	7.5	8.0

Det ses, at der er flere middelværdier, der bliver identiske. Eksempelvis er der hele 8 middelværdier, der er lig med 4.5. Der kan nu opstilles en frekvenstabel over middelværdierne, på den måde, som det blev gjort i noten om *deskriptiv statistik*. Dette er gjort i følgende tabel, hvor middelværdi, variansen og standardafvigelsen tillige beregnes. Dette er gjort på samme måde, som det var tilfældet for totalpopulationen.

X	**P(X)**	**XP(X)**	**(X-μ)**	**$(X-\mu)^2$**	**$P(X)(X-\mu)^2$**
1.0	0.0156	0.0156	-3.5	12.25	0.1914
1.5	0.0313	0.0469	-3.0	9.00	0.2813
2.0	0.0469	0.0938	-2.5	6.25	0.2930
2.5	0.0625	0.1563	-2.0	4.00	0.2500
3.0	0.0781	0.2344	-1.5	2.25	0.1758
3.5	0.0938	0.3281	-1.0	1.00	0.0938
4.0	0.1094	0.4375	-0.5	0.25	0.0273
4.5	0.1250	0.5625	0.0	0.00	0.0000
5.0	0.1094	0.5469	0.5	0.25	0.0273
5.5	0.0938	0.5156	1.0	1.00	0.0938
6.0	0.0781	0.4688	1.5	2.25	0.1758
6.5	0.0625	0.4063	2.0	4.00	0.2500
7.0	0.0469	0.3281	2.5	6.25	0.2930
7.5	0.0313	0.2344	3.0	9.00	0.2813
8.0	0.0156	0.1250	3.5	12.25	0.1914
Sum	1.0000	4.5000			2.6250

Følgende værdier fås:

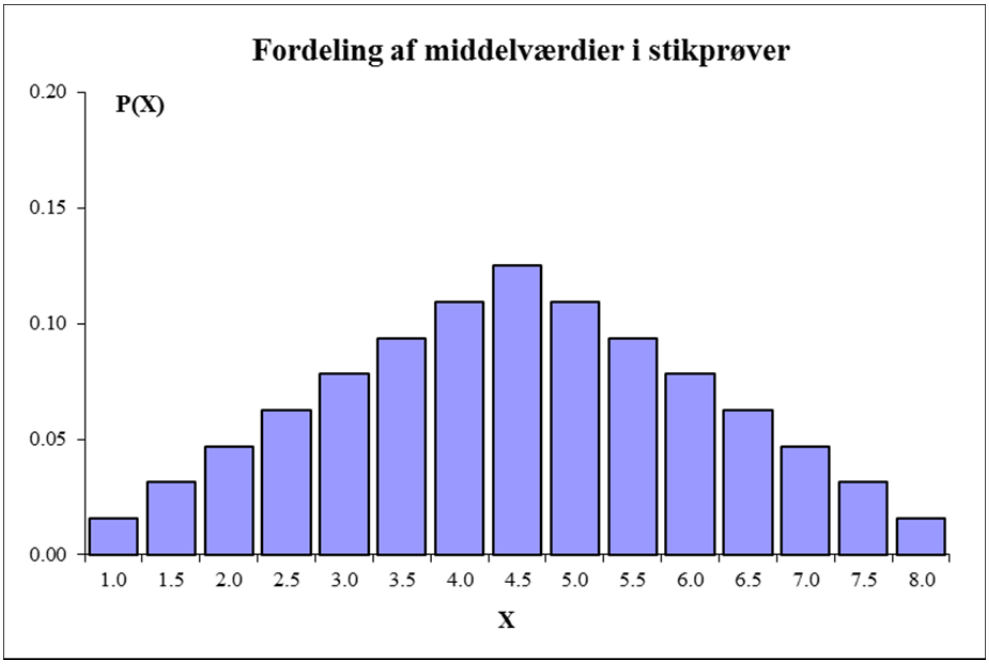
$$E(\overline{X}) = \mu = 4.5 \qquad V(\overline{X}) = s_{\overline{X}}^2 = 2.6250 \qquad SD(\overline{X}) = s_{\overline{X}} = 1.6202$$

Dette er overraskende! Mens middelværdien mellem totalpopulationen og stikprøven er identiske, så er *variansen blevet mindre*. Det kan helt præcis udregnes, hvor meget variansen er reduceret. Divideres variansen fra totalpopulationen, der var lig med 5.25, med

størrelsen af stikprøven, der var lig med 2, så får man 2.6250, som er variansen i stikprøven. Det vil sige, at der gælder følgende sammenhænge:

$$E(\bar{X}) = \mu \qquad V(X) = s_X^2 = \sigma^2 / n \qquad SD(\bar{X}) = s_X = \sigma / \sqrt{n}$$

Betragt nu fordelingen af *X* og *P(X)* i diagramform, hvor data er taget fra første og anden kolonne i tabellen ovenfor:



Denne "pyramide" ligner Normalfordelingen. Man kan opsummere den foregående analyse som følger:

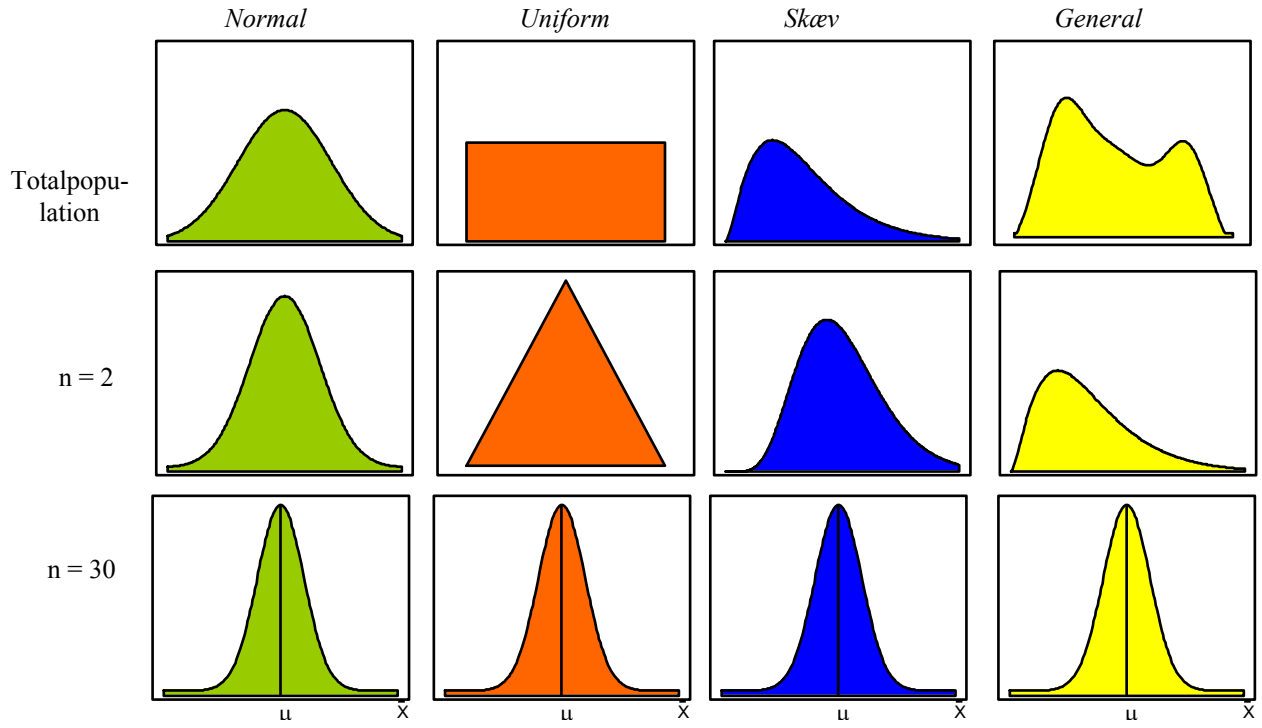
Når man tager en stikprøve fra en uniformt fordelt totalpopulation, så vil enhver stikprøve være Normalfordelt med middelværdi og varians givet som:

$$X \approx U(\mu, \sigma^2) \rightarrow X \approx N(\mu, \frac{\sigma^2}{n})$$

Dette udtryk kan generaliseres til at være gældende, uanset hvilken fordeling som totalpopulationen følger. Middelværdien forbliver den samme, mens variansen mindskes med en faktor 1/n eller for standardafvigelsen med faktoren $1/\sqrt{n}$. Det vil sige, at estimatet på middelværdien er en *unbiased* estimator af μ. Endvidere gælder, at den første af de tre ovennævnte karakteristika er opfyldt.

Hvad sker der, når størrelsen af stikprøven øges? Dette vises i illustrationen øverst på næste side.

Konsekvenser af den central grænseværdisætning: Fordelingen af $\overline{X}$ for stigende størrelse af stikprøven



Uanset hvilken fordeling, som totalpopulationen har, så bliver stikprøven normalfordelt. Når stikprøvens størrelse øges, så sker der yderligere det, at variansen i stikprøven reduceres med faktor n i forhold til variansen i totalpopulationen. Det betyder, at de to karakteristika om efficiens og konsistens er opfyldt.

Eksemplet leder til den centrale grænseværdisætning, som kan formuleres som i rammen:

Den centrale grænseværdisætning

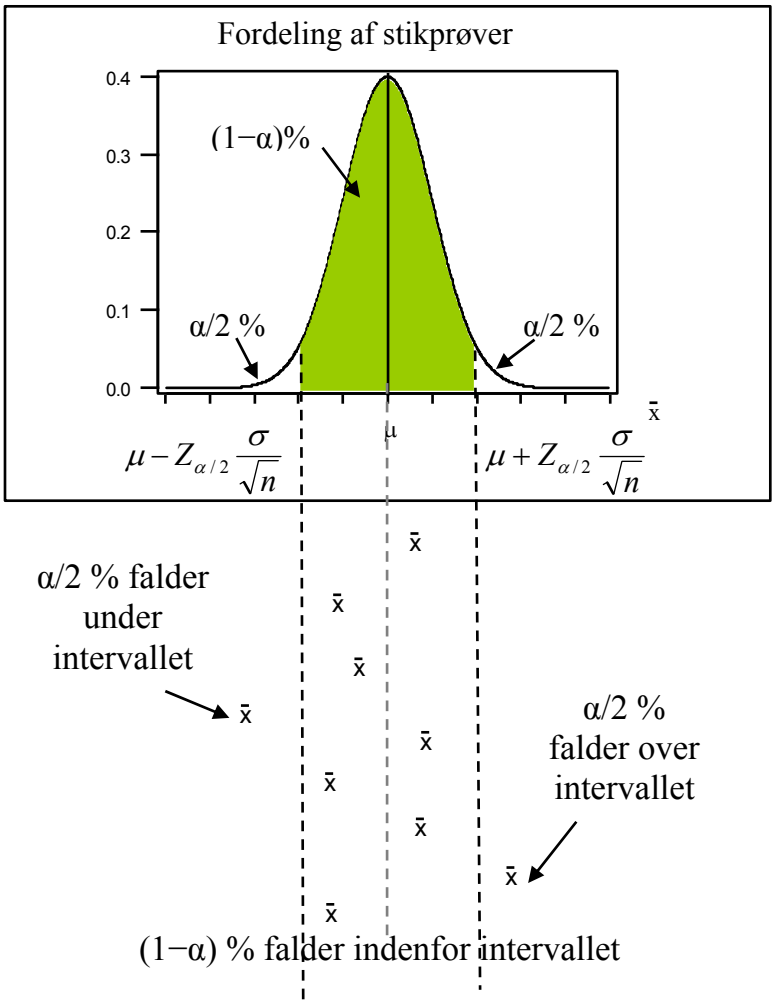
Ved udtagning af en stikprøve fra en totalpopulation med middelværdi lig μ og standardafvigelse lig σ, vil fordelingen af stikprøvens middelværdi $\overline{X}$ tendere at følge en Normalfordeling med middelværdi lig μ og standardafvigelse lig $\sigma/\sqrt{n}$ når størrelsen af stikprøven øges

For "tilstrækkelig stor" n gælder at $\overline{X} \approx N(\mu, \sigma^2/n)$

Dette er et vigtigt resultat! Hvorfor? Jo – nu kan man generelt antage, at alle stikprøver er Normalfordelte eller Binomialfordelte. Sidstnævnte kunne jo omformuleres til en Normalfordeling. Dette har betydning, når der i det følgende arbejdes med betragtninger omkring usikkerhed.

2. Konfidensinterval for middelværdien

I eksemplet i det sidste afsnit blev det vist, at fordelingen af mulige stikprøver i en totalpopulation vil følge en Normalfordeling. Denne information kan anvendes til at opstille et usikkerhedsmål for spredningen omkring middelværdien eller populationsandelen. Et *konfidensinterval* angiver de grænser, inden for hvilke man er sikker på, at den estimerede parameter (μ eller $\hat{p}$) befinder sig.



Tankegangen er illustreret ovenfor. Forskellige stikprøver fra den samme totalpopulation vil resultere i forskellige middelværdier. Ganske som i eksemplet. Derfor søges et usikkerhedsmål for, hvor langt væk fra den mest hyppige middelværdi, man kan tillade at være, uden at man siger, at stikprøven er skæv eller biased.

Hvis stikprøven ikke er skæv, så vil middelværdien i stikprøven $\overline{X}$ tilnærmelsesvis være lig med middelværdien i totalpopulationen μ. Fordelingen af middelværdierne følger Normalfordelingen, der er tabuleret i **Statistics Tables** og benævnes *Z*. Usikkerheden i fordelingen er variansen eller standardafvigelsen.

I notesættet om Normalfordelingen blev den inverse sammenhæng mellem en stokastisk variabel X og middelværdien og variansen givet som $X = \mu_x + Z\sigma_x$. Vores interesse er ikke as se på en given værdi, men på en tolerance i forhold til middelværdien. Så i vort tilfælde er $X = 0$. Lad denne tolerance været givet ved usikkerheden **α**. Tolerancen kan, som det er illustreret i figuren ovenfor, være såvel positiv som negativ. Da er usikkerheden **α/2.** Erstattes lighedstegnet med "±" og ved lidt ombytning fås:

$$X \pm Z_{\alpha/2}\sigma_x$$

Dette udtryk anvendes, hvis størrelsen af stikprøven ikke er kendt.

Kendes størrelsen af stikprøven anvendes følgende udtryk, idet man anvender spredningen, som fundet i den centrale grænseværdisætning:

$$\overline{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Formlen angiver beregningen af et (1−α)×100 interval. Inden for dette interval vil man kunne forvente at finde hovedparten af de mulige middelværdier for stikprøverne.

Tolerancerne for konfidensintervallet afhænger af *konfidensnivauet.* Det typiske konfidensniveau er på *95 procent.* Det vil sige, at α = 0.05 og (1−0.05)×100 = 95 procent. Da intervallet er tosidet, så vil den nedre grænse ligge ved 0.05/2 svarende til 0.025, og den øvre grænse ved (1 – α/2) svarende til 0.975.

Nu findes værdierne af Z, som svarer til disse sandsynligheder i **Statistics Tables**. Man behøver kun at finde den ene af værdierne. Dette skyldes, at Normalfordelingen er symmetrisk omkring middelværdien nul jævnfor det forrige sæt af noter.

Z	**0.00**	**…**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
1.6	0.9452	…	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	…	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	…	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	…	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	…	0.9798	0.9803	0.9808	0.9812	0.9817

Her findes der for den øvre grænse at Z = 1.96. Find, som en øvelse, den tilsvarende negative værdi for 0.025. For et 95 procents konfidensinterval haves således Z = ±1.96.

De mest almindelige konfidensintervaller er på 90, 95 eller 99 procent. *Hvis der ikke står noget opgivet i en opgave, så anvendes altid et 95 procents konfidensinterval.* De tilhørende værdier af Z for de forskellige konfidensintervaller er givet som:

Konfidensinterval:	Værdi af Z
90 (α = 0.10) 95 (α = 0.05) 99 (α = 0.01)	± 1.645 ± 1.960 ± 2.575

Find, som en øvelse, Z-værdierne for 90 og 99 procents konfidensintervallerne.

Eksempel
Antag en stikprøve/datasæt med 100 observationer og middelværdi lig med 50. Endelig er standardafvigelsen lig med 5. Opstil et 90, 95 og 99 procents konfidensinterval for middelværdien.

For at løse problemstillingen anvendes formlen ovenfor og de fundne værdier for Z. Følgende fremkommer:

90 %: $\overline{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \Rightarrow 50 \pm 1.645\frac{5}{\sqrt{100}} \Rightarrow 50 \pm 0.8225 \Rightarrow [49.1775\,;50.8225]$

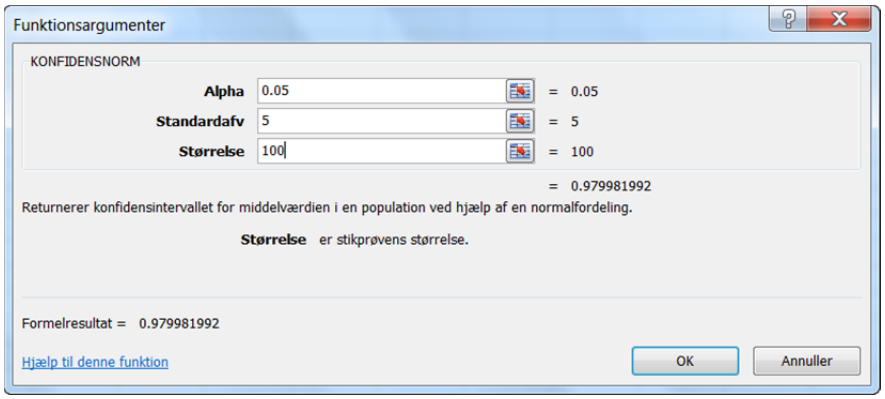
95 %: $\overline{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \Rightarrow 50 \pm 1.960\frac{5}{\sqrt{100}} \Rightarrow 50 \pm 0.9800 \Rightarrow [49.0200\,;50.9800]$

99 %: $\overline{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \Rightarrow 50 \pm 2.575\frac{5}{\sqrt{100}} \Rightarrow 50 \pm 1.2875 \Rightarrow [48.7125\,;51.2875]$

Bemærk konfidensintervallet øges med stigende konfidensniveau.

Konfidensintervallet kan også beregnes ved anvendelse af **TI-84 lommeregneren.** Tryk STAT → TESTS → 7: Zinterval. I den anden linje skal STATS være markeret. Nu indsættes værdierne i menuen. I dette tilfælde er σ = 5, $\overline{X}$ = 50 og n = 100. Vælg C-level: .95 → CALCULATE → ENTER. Da fremkommer resultatet som [49.02;50.98] ganske som ovenfor.

Endelig kan man anvende **Excel.** Her brug **Formler | Indsæt funktion | statistisk | konfidensnorm.** Da fremkommer nedenstående dialogboks.



Bemærk, at man selv skal beregne den øvre og nedre grænse ud fra de værdier, som Excel returnerer.

Eksempel: Et konfidensinterval for prisen på benzin
I det forrige sæt af noter, så man på prisen på benzin ud fra en forudsætning om, at prisen var Normalfordelt. Prisen på 95 oktan var over 25 besøg på tankstationen fundet til at være lig med 9.95 DKK med en standardafvigelse på 0.30 DKK.

Nu opstilles et 95 procents konfidensinterval for prisen:

$$\overline{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \Rightarrow 9.95 \pm Z_{0.025}\frac{0.30}{\sqrt{25}} \Rightarrow 9.95 \pm 1.96(0.06) \Rightarrow 9.95 \pm 0.1176 \Rightarrow \left[9.8324\,;10.0676\right]$$

Fortolkningen er, at i 95 procent af de gange, hvor der aflægges et besøg på tankstationen, vil prisen variere inden for dette interval. Det vil sige en pris mellem 9.83 DKK og 10.07 DKK.

Eksempel med simpel undersøgelse af en påstand (eksamen juli 2012, 10 %, 2P)
Meget ofte anvendes et konfidensinterval til at undersøge validiteten af en påstand. Der opstilles først et konfidensinterval for et datasæt. Dernæst sammenholdes påstanden med konfidensintervallet. Falder påstanden udenfor intervallet siges denne at være afvist, mens påstanden er accepteret, hvis denne falder inden for intervallet.

Problemstilling
Togturen fra Hamburg til Flensborg med "Regionalexpress"-forbindelsen tager præcis 120 minutter. 100 repræsentative målinger af den samme strækning med bil ad motorvejen resulterer i en middelværdi på 110 minutter ved en standardafvigelse på 30 minutter (pga. hyppigt opstående kødannelser). Adskiller bilen sig signifikant fra de 120 minutter med toget mht. hastigheden?

Løsning
Først skal der findes frem til *påstanden.* Det må være en sammenligning mellem de 120 minutter, som togturen tager i forhold til turen med bil.

Umiddelbart ser turen med bil ud til at være 10 minutter hurtigere, men der kan være kø, vejarbejde, fodbold på Volkspark eller hvad ved jeg! Det giver en usikkerhed.

Om kørsel med bil vides det, at der er foretaget tidsmåling på 100 ture under alle former for trafiktæthed. Det vil sige at n = 100. Endvidere er der opgivet middelværdi og standardafvigelse. Det vigtigste i formuleringen af opgaven er, at der står opgivet, at materialet er *normalfordelt.* Det er et hint til, at der skal anvendes et konfidensinterval.

Der indsættes nu for de givne informationer:

$$\overline{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \Rightarrow 110 \pm Z_{0.025}\frac{30}{\sqrt{100}} \Rightarrow 110 \pm 1.96(3) \Rightarrow 110 \pm 5.88 \Rightarrow [104.12\,;115.88]$$

Det ses, at de 120 minutter ligger ovenfor intervallet. Det vil sige, at kørsel med bil i 95 procent af tilfældene er det hurtigste.

4. Konfidensinterval for populationsandelen

Som det er tilfældet med middelværdien kan der opstilles et konfidensinterval for populationsandelen *p*. Det kan eksempelvis være et konfidensinterval for en andel af stemmer eller for markedsandelen af et givet produkt.

Stikprøveandelen $\hat{p} = \frac{x}{n}$ er estimator for populationsandelen *p* i totalen. Her angiver *x* antallet af for eksempel individer, der stemmer "ja", mens *n* er antallet af observationer.

Hvad er middelværdien og standardafvigelsen for populationsandelen? Når der arbejdes med populationsandele, så er der kun to udfald, nemlig succes *p* eller fiasko *q*, som er lig med *(1−p)*. Populationsandelen er derfor en Binomialfordelt stokastisk variabel. I det sidste sæt af noter blev det i det afsluttende afsnit vist, hvordan en Binomialfordelt stokastisk variabel kan tilnærmes en Normalfordelt stokastisk variabel for tilstrækkeligt store værdier af stikprøven *n*. Denne viden udnyttes nu, idet der indsættes for middelværdi og standardafvigelse i det i sidste afsnit udledte udtryk for konfidensinterval for middelværdien.

Så for en tilstrækkelig stor størrelse af stikprøve *n* vil middelværdien af andelen for et enkelt forsøg være lig med $\hat{p}$. Variansen for et enkelt forsøg vil være lig med $\hat{p}(1-\hat{p})$. Se også notesæt 2 side 12. Udtrykket $s/\sqrt{n}$ kan ved indsættelse omformuleres til at gælde for andele. Her fås $\sqrt{\hat{p}(1-\hat{p})/n}$. I alt haves da:

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Formlen angiver beregningen af et (1−α)×100 interval for populationsandelen. Inden for dette interval vil man kunne forvente at finde hovedparten af de mulige populationsandele for stikprøverne.

Hvad er en tilstrækkeligt stor stikprøve? Her kan der opstilles følgende regler: Stikprøven er stor når der gælder, at såvel $n \cdot \hat{p}$ og $n \cdot (1 - \hat{p})$ er mindst lig med 5.

Eksempel: Bageren på Langeland
Igennem de seneste årtier er der blevet færre kommuner i Danmark. Eksempelvis blev de fem kommuner på Bornholm i år 1999 efter en afstemning sammenlagt til en enkelt kommune.

Ved kommunevalget den 12. februar 2003 blev der foretaget en lignende afstemning på øen Langeland i det fynske øhav. Indtil da var der tre kommuner på Langeland, som via en afstemning ønskede at blevet lagt sammen til en enkelt kommune.

En lokal bager i hovedbyen Rudkøbing havde op til valget lavet en interessant form for prognose på udfaldet af afstemningen. På de Fastelavnsboller, som bageren solgte, kunne kunden få skrevet "ja" eller "nej" i chokolade afhængigt af, hvilken overbevisning, som kunde havde.

Op til datoen for afstemningen solgte bageren i alt 385 Fastelavnsboller dekoreret med "ja", mens 271 dekoreret med et "nej".

- Opstil et 95 procents konfidensinterval for populationsandelen af kager med påtrykt "ja" for sammenlægning af kommunerne

Det samlede antal solgte Fastelavnsboller er lig med $n = 385 + 271 = 656$.

Andelen af Fastelavnsboller med "ja" er lig med $\hat{p} = x / n = 385 / 656 = 0.587$. Nu anvendes formlen ovenfor, da det vides at $Z_{0.025} = 1.96$

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \Leftrightarrow 0.587 \pm 1.96\sqrt{\frac{0.587(1-0.587)}{656}} \Leftrightarrow 0.587 \pm 1.96(0.019) \Leftrightarrow 0.587 \pm 0.038$$

Et 95 procents konfidensinterval er da lig med $[0.549\,;0.625]$.

Betingelsen $n \cdot \hat{p} \geq 5$ samt $n \cdot (1 - \hat{p}) \geq 5$ er begge opfyldt.

Eksempel med simpel undersøgelse af en påstand (eksamen juli 2012, 10 %, 2P)
Som i det forrige afsnit kan man anvende et konfidensinterval for en populationsandel til at undersøge validiteten af en påstand. Fremgangsmåden er som før. Den eneste forskel er, at der nu arbejdes med andele.

Problemstilling
Et motionscenter i Flensborg har en markedsandel på 20 %. Centeret gennemfører en reklamekampagne. Derefter spørges 250 repræsentativt udvalgte personer, om de ønsker at komme i dette motionscenter, og 25 % siger ja. Motionscentrets ledelse vurderer, at markedsandelen er steget som følge af reklamekampagnen. Kan ledelsens antagelse underbygges statistisk?

Løsning
Som i det forrige afsnit, skal der først findes frem til *påstanden.* Motionscentret har en andel på 20 % og dernæst spørges 250 personer, hvor de 25 % siger at de bruger centret.

Det må være en sammenligning mellem de 20 %, som har svaret først og de 250 personer, hvor 25 % svarer at de bruger centret. Det vil side at $n = 250$ og $\hat{p} = 0.25$.

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \Leftrightarrow 0.25 \pm 1.96\sqrt{\frac{0.25(1-0.25)}{250}} \Leftrightarrow 0.25 \pm 1.96(0.027) \Leftrightarrow 0.25 \pm 0.053$$

Et 95 procents konfidensinterval er da lig med $[0.197\,;0.303]$.

Betingelsen $n \cdot \hat{p} \geq 5$ samt $n \cdot (1-\hat{p}) \geq 5$ er begge opfyldt.

Da de 20 % svarende til 0.20 netop ligger inden for intervallet gælder påstanden ikke! Det vil sige, at med det givne materiale og usikkerhed er de 20 % ikke signifikant forskellige fra de 25 %.

Uddybning (ikke del af eksamensspørgsmål)
To forhold kan diskuteres. Der kan argumenteres for, at anvende de 0.20 i stedet for 0.25, da opgaven er uklart formuleret. Her opnås den samme konklusion. Konfidensintervallet bliver med $[0.1504\,;0.2495]$, og de 0.25 ligger *lige over* intervallet.

Noget tyder på, at der er blevet spurgt for få personer i undersøgelsen. Var antallet af respondenter for eksempel 300 ville divisoren være blevet større, og konfidensintervallet mindre. Derved ville undersøgelsens *efficiens* være blevet øget, og udfaldet ville være blevet mere *entydigt.*

Nils Karl Sørensen ©
Statistik I, BA int, Flensborg

Sæt 4: Korrelation og kovarians

af Nils Karl Sørensen

Indhold ***side***

1. Korrelation og kovarians

Da vi arbejdede med *deskriptiv statistik,* så vi blandt på sammenhængen mellem to variable kaldet x og y. Ved at konstruere et krydsdiagram for n par af observationer (x_i,y_i) for alle observationer $i = 1,2, \ldots n$ kan der opnås indsigt i, om sammenhængen er positiv-, negativ eller ikke eksisterende.

Hvis en positiv eller negativ sammenhæng eksisterer, ønsker vi at opstille et mål for, hvor god den lineære sammenhæng er. *Kovariansen* er et udtryk for denne samvariation. Den benævnes S_{xy} og defineres som:

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

For at anvende dette udtryk skal man først finde middelværdierne for x og y. Derefter tages for hvert par af observationer forskellen til middelværdien. Disse to størrelser multipliceres med hinanden. Afslutningsvis summeres, og der divideres med $(n-1)$.

Kovariansen kan være positiv, negativ eller nul. I det sidste tilfælde findes der ingen relation mellem de to variabler.

Et problem med at anvende kovariansen i den skitserede form er, at størrelsen afhænger af de enheder, som x og y måles i. Hvis man eksempelvis måler i DKK, så vil kovariansen være meget mindre end det er tilfældet, hvis man måler i 1,000 DKK.

Dette problem kan man løse ved at normalisere kovariansen i forhold til de data, der anvendes.

Korrelationskoefficienten er et udtryk for styrken af samvariationen mellem to variabler. Her normaliseres med stikprøve standardafvigelserne. Korrelationskoefficienten defineres som:

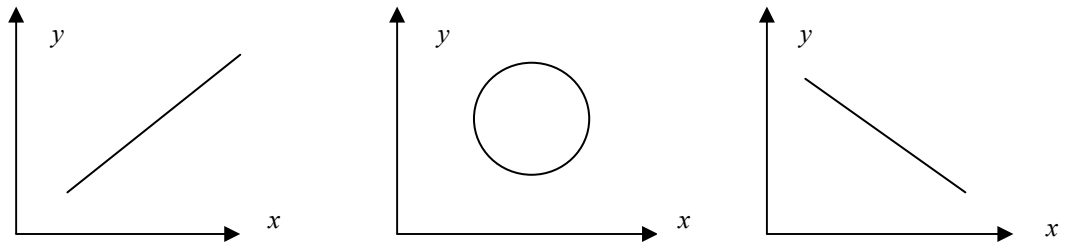
$$r = \frac{S_{xy}}{S_x S_y}$$

Hvor S_x og S_y er stikprøve standardafvigelserne af henholdsvis x og y. De er defineret som:

$$S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \quad \text{og} \quad S_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

Defineret på denne måde vil variationen af korrelationskoefficienten være lig $-1 \le r \le 1$. Desto nærmere korrelationskoefficienten er på ± 1, desto stærkere er korrelationen. Hvis korrelationen er tæt på nul, så er der ingen sammenhæng mellem x og y.

Eksempler på korrelationer



Positiv **Ingen sammenhæng** **Negativ**

Korrelationen anvendes ofte til at beregne *determinationskoefficienten*, som er givet ved:

$$R^2 = r^2$$

R^2 varierer som $0 \le R^2 \le 1$. Den er en slags procentuelt mål for styrken af sammenhængen. Hvis R^2 er lav, så er sammenhængen lav, da variablen x ikke forklarer, så meget af variation i y. Omvendt hvis R^2 er høj. Der vendes tilbage til R^2 i notesæt 5 til Statistik II om *regression.*

2. Eksempel *(Eksamen februar 2012, Opgave 3. (15 %, 3P))*

Er en god rating af filmanmelderne en garanti for, at mange folk også ser en film i biografen?

Filmmagasinet "Ekko" har undersøgt dette i sin decemberudgave 2011. Tabellen viser antal solgte billetter X og ratings Y i fire toneangivende aviser for 24 film fra 2002 til 2011, som henvender sig specifikt til teenagere.

År	Film	X: Billetter (1000)	Y: Stjerner
2002	Slim, slam, slum	7.8	1.6
2003	Midsommer	121.3	3.4
	2 ryk og en aflevering	78.2	3.6
	Bagland	90.6	4.3
2004	Tæl til 100	23.1	2.4
	Terkel i knibe	375.8	4.4
2005	Af banen	71.7	2.8
	Unge Andersen	10.0	3.3
	Strings	15.7	3.8
2006	Supervoksen	106.5	3.6
2007	Fighter	52.2	3.5
	Rich Kids	107.0	2.6
2008	To verdener	314.5	4.0
	Dog og mig	101.8	3.6
	Rejen til Saturn	401.0	4.0
2009	Kærestesorger	175.1	3.6
	Se min kjole	45.4	4.0
	Vanvittig forelsket	27.4	4.0
2010	Hold om mig	46.0	3.2
	Min bedste fjende	9.4	3.4
2011	Frit fald	29.2	4.6
	Bora bora	75.4	2.8
	Skyskraber	1.1	3.6
	Magi i luften	3.0	3.3
Sum		2289.2	83.4

Desuden er de efterfølgende karakteristiske værdier allerede beregnet:

$$\bar{x} = 95.38, \quad \bar{y} = 3.48, \quad \sum_{i=1}^{24} x_i^2 = 514159.64, \quad \sum_{i=1}^{24} y_i^2 = 300.4; \quad \sum_{i=1}^{24} (x_i - \bar{x})(y_i - \bar{y}) = 676.22$$

A. Beregn standardafvigelserne for X og Y. **1P**

B. Beregn et mål for sammenhængen mellem X og Y og fortolk resultatet. **2P**

Løsning
Opgaven kan også beregnes på en **lommeregner.** På **TI-84eren** tastes ind som henholdsvis *L1* og *L2*. For hver series fås standardafvigelsen ved at beregne: STAT → CALC → 1: 1-Var Stats L1, og tilsvarende en gang til for L2.

Korrelationen findes ved at beregne STAT → CALC → 4: LinReg(ax+b) *L1,L2* → r (findes i den sidste linje i outputtet i skærmbilledet på lommeregneren)[2]

A)
For at beregne standardafvigelserne kan man bruge den beregningsformel, der er anvendt i en lignende opgave tidligere. Se også notesæt 1 side 24 nederst. Summen af værdierne kan beregnes på 2 måder. Enten ved at tage gennemsnittet og gange med antallet af observationer (her fås et lidt andet resultat for y-værdierne, da der er flere decimaler at regne med). Alternativt kan man lægge de 24 observationer sammen. Man får:

$$s_x = \sqrt{\frac{1}{24-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right)} = \sqrt{\frac{1}{24-1}\left(514159.64 - \frac{(2289.2)^2}{24}\right)} = 113.41$$

$$s_y = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}\right)} = \sqrt{\frac{1}{24-1}\left(300.4 - \frac{(83{,}4)^2}{24}\right)} = 0.68$$

B)

Korrelationen giver bedst denne sammenhæng. Først beregnes kovariansen:

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{676.22}{24-1} = 29.40$$

Korrelationen kan nu beregnes som:

[2] Denne funktion anvendes også til simpel regression. Der vendes som nævnt tilbage til dette emne i Statistik II

$$r = \frac{S_{xy}}{S_x S_y} = \frac{29.40}{113.41 \times 0.68} = 0.3812$$

Der fremkommer som ventet en positiv korrelation, som indikerer den postulerede positive sammenhæng.

Nils Karl Sørensen ©
Statistik I, BA int, Flensborg

Sæt 5: Skalanivauer og krydstabeller

af Nils Karl Sørensen

Indhold *side*

1. Indledning

Indtil nu har vi i kurset primært arbejdet med undersøgelser af enkelte variabler opgjort på forskellig måde eksempelvis som gennemsnit. Nu vil vi se på, hvilke typer af variabler, der findes, og på hvordan de opgøres. Dernæst vil vi, som en fortsættelse af korrelations analysen, se på, hvordan man tabeullerer to sæt at samhørende variabler.

2. Skalaniveauer

Når man foretager en optælling til for eksempel en tabel, benævnes det, som man tæller og klassificerer, *elementer* eller observationer. Det kan være kunder, biler eller antallet af studerende på en given uddannelse.
Klassifikation består i en opdeling af elementerne i kategorier efter bestemte kriterier eller kendetegn. Der findes to typer af kriterier, som man kan klassificere sine observationer efter; de direkte målbare og de indirekte målbare. Lad os se på disse efter tur.
Direkte målbare kriterier benævnes ***kvantitative kriterier***. De konstateres ved, at der en numerisk værdi tilordnet tællingen. Det kan for eksempel være priser opgjort i € eller DKK eller opgjort som et indeks. Andre eksempler er måling af en højde i centmeter eller vejning i kilo af en persons vægt. Kvantitavie data har vi anvendt til at beregne gennemsnit, spredning og andre mål. Vi har anvendt de beregnede værdier til for eksempel at fastlægge udseendet af fordelingen af det analyserede materiale.
Indirekte eller *ikke målbare kriterier* benævnes ***kvalitative kriterier***. De konstateres med karakteristika, som eksempelvis køn, civilstand, bopæl, fabrikat af en bil eller erhvervs- og uddannelsesniveau. Tællingsenhederne kan sorteres og optælles efter de kvalitative kriterier, men man kan ikke tillægge køn eller erhverv en bestemt værdi, hvorfor der ikke kan beregnes eksempelvis et gennemsnit af disse. Derimod kan man optælle sit materiale på køn, og dernæst for eksempel beregne den gennemsnitlige indkomst for henholdsvis kvinder og mænd. Ved en sådan beregning kan man hensigtsmæssigt tildele en talkode til en

variabel. For eksempel kan man sige, at kvinde=1 og mand=0. Dette letter arbejdet med at sortere, men tallet er ikke i sig selv udtryk for et kvantitativt mål.
De kvantiative og de kvalitative kan hver især opdeles i to grupper afhængig af klassificeringen. Følgende 4 klassificeringsmåder (målskalaer) fremkommer derved:

Kvalitative kriterier:	1) Uegentlig skala (nominalskala)	
	2) Ordinalskala	
Kvantitative kriterier:	3) Referencepunktskala	
	4) Nulpunktspunkskala	Intervalskalaer

***Ad 1) Uegentlig skala (nominalskala)*:** Klassifikation af elementer efter kvalitative kriterier uden en bestemt orden.
Klassificeringen består her alene af en sortering. Elementerne er eksempelvis personer med fast bopæl i Danmark, som kan opdeles (sorteres) efter køn, erhverv eller bopælskommune. Rækkefølgen er uden betydning.
Ad 2) Ordinalskala: Her er rækkefølgen ikke vilkårlig. Det kan eksempelvis være en opdeling af lande efter kreditværdighed (AAA, AA+), eller det kan være nytteniveauer, som er kendt fra mikroøkonomien.
Et andet eksempel findes typisk i spørgeskemaundersøgelser, hvor man kan have forskellige typer af udfald af svar, der krydses af. Det kan se ud som følger:

☐	☐	☐	☐	☐
Meget tilfreds	Tilfreds	Neutral	Utilfreds	Meget utilfreds

For begge eksempler gælder, at man ikke kan sige, hvor stor en forskel, der er på eksempelvis AAA eller AA+, eller på svarene ”tilfreds” eller ”neutral”. Yderligere vil vurderingen af udsagnet være individuel. To forskellige kreditvurderingsselskaber kan nå frem til forskellige vurderinger, og to forskellige respondenter kan have forskellige opfattelser af, hvornår man er ”tilfreds” eller ”neutral”.
Imidlertid kan kan man godt foretage en form for beregninger på ovennævnte materiale, idet man tilordner et tal til udfaldet. Eksempelvis kan man tænke sig, at svarmulighederne ovenfor til ordnes som følger:

☐	☐	☐	☐	☐
Meget tilfreds=1	Tilfreds=2	Neutral=3	Utilfreds=4	Meget utilfreds=5

Nu kan der foretages visse beregninger på materialet. Antag at der er 200 respondenter i undersøgelsen, og deres svar fordeler sig som følger:

Svar	Meget tilfreds	Tilfreds	Neutral	Utilfreds	Meget utlfreds	
Talværdi	1	2	3	4	5	
Frekvens	20	40	80	30	30	200

Der kan nu beregnes et gennemsnit og andre lokations- og spredningsmål. Vi får eksempelvis at gennemsnittet er 3.05 og standardafvigelsen er lig 1.16. Det betyder, at det

typiske svar er lig med 3 = ”neutral”. Endvidere falder hovedparten af svarene indenfor kategorierne 2 til 4. Der kan tilsvarende laves en grafisk præsentation af materialet.
Lad os nu antage, at der er stillet et andet spørgsmål, hvor der er opnået følgende svar:

Svar	Meget enig	Enig	Neutral	Uenig	Meget uenig	
Talværdi	1	2	3	4	5	
Frekvens	60	30	20	40	50	200

Kan man sammenligne de 2 fordelinger opgjort på en ordinalskala? Umiddelbart er svaret *nej*. Årsagen er, at ”tilfreds” ikke nøvendigvis er det samme somat være ”enig”. Er gennemsnittet det samme for de to datasæt, kan der heller ikke foretages en sammenligning. Det er således et krav for at kunne foretage en sammenligning, at der er konsistens mellem de udfaldene i det, som der spørges om. Ordinalskalaen er således vanskelig at arbejde med.
Lad os nu gå til de ***kvantitative kriterier***. Imodsætning til eksemplet ovenfor, så kan måleskalaen her tillægges en betydning. En ***intervalskala*** er således en måleskala for kvantitative data. Intervalskalaen kan være enten i *punktform* eller i *intervalform.* Vi har set på begge skalaer i forbindelse med den diskriptive statistik. I *punktform* betragter vi for eksempel en tidsserie som 1,2,3 osv., mens en intervalform opdeler i en række intervaller eksempelvis [0−9];[10−19] osv. Bemærk at udgangspunktet er forskelligt. I dette eksempel starter den ene af vores klassificeringer i punktet 1, mens den anden tager begyndelsen i punktet 0. Hvad er det korrekte? For at kunne bestemme dette, må vi se på de 2 typer af skalaer for de kvantitative klassifikationer.
Ad 3) Referencepunktskala: Klassifikation af elementer efter kvantitative kriterier og inddelt efter en intervalskala med vilkårligt valgt referencepunkt.
Eksempler er Celsius- og Fahrenheit-temperaturskalaerne. Tidsaksen er et andet eksempel. I den europæiske kulturkreds refererer vore årstal til Kristi fødsel. Japanerne refererer til året for den siddende kejsers tronbestigelse, mens muslimerne refererer til til år 622 e. Kr. Andre eksempler er karakterskalaen (såvel den tyske som den danske).
Ad 4) Nulpunksskala: Klassifikation af elementer efter kvantitative kriterier og inddelt efter en intervalskala med absolut eller entydigt nulpunkt.
Antal, vægt, højde, længde, afstand, diameter, areal, rumfang, pris, indtægter, udgivfter, indkomst, produktion, forbrug, opsparing, investering, alder og hastighed er alle eksempler på kendetegn, hvis størrelse fastsættes ud fra en nulpunktsskala.

Eksempel
Opgave 7, sommereksamen 2011. (3 points, 15 %)
En biologisk undersøgelse beskæftiger sig med antallet af eksisterende sommerfuglearter i forskellige biotoper i den tyske delstat Schleswig-Holstein. I den forbindelse undersøgte man forskellige skove samt eng- og græsområder i samme antal og bestemte de eksisterende sommerfuglearter. Basislisten på næste side viser følgende værdier for de eksisterende sommerfuglearter i de forskellige biotoper.

Område	Eng/græs
1	8
2	11
3	10
4	10
5	8
6	14
7	11
8	14
9	11

Område	Skov
1	8
2	10
3	5
4	5
5	5
6	12
7	17
8	5
9	8

A) **0,5P**
Hvilke(t) skalaniveau(er) foreligger der?

Måleenheden er ”antallet af sommerfuglearter” opgjort efter ”biotoper”. Det er levestedet. Dvs. at henholdsvis 2. og 4. søjle er kvantitative data opgjort efter område. Biotopen er en *ordinal variabel*.
Forspalten er imidlertid gengivet med et referencenummer til et givet område, som kan være gengivet helt tilfældigt. Det må være en *Uegentlig skala (nominalskala)*, da der ikke fremgår nogen bestemt orden.

B) **1,5P**
Beregn alle mulige lokationsmål antallet af sommerfugle i begge biotoper i overensstemmelse med skalaniveauet.

Lokationsmål er: Middeltallet, medianen og typetallet (den hyppigste observation). Eventuelt kan øvre og nedre kvartil medtages.
Lokationsmålene beregnes med **lommeregneren,** som beskrevet i notesæt 1. Den lille tabel kan opsummere resultaterne:

Mål	Eng/græs	Skov
Gennemsnit	10,78	8,33
Typetallet	11	5
Median	11	8
Q_1	9	5
Q_3	12,5	11

Typetallet har jeg fundet ved manuel granskning af de 2 tabeller!

C) **0,5P**
Hvilken fordeling foreligger for begge biotoper? Begrund dit svar

Gennemsnit, median og typetal er meget identiske for fordelingen af sommerfuglearter på "eng/græs". Dvs. at fordelingen er *symmetrisk*. Typetallet for sommerfuglearter i skoven er mindre end median og gennemsnit. Fordelingen er således *højreskæv.*

D) **0,5P**
Hvis du på en søndagsudflugt skulle iagttage sommerfugle, hvilken biotoptype ville du så opsøge? Begrund dit valg, idet du benytter dig af de tidligere opnåede resultater.

Jeg ville vælge "eng/græs", da man her finder det største antal arter, og den mindste spredning fordelt på områder.

3. Krydstabeller

Tabellen er det vigtigste værktøj inden for praktisk statistik. I koncentrerede, systematiske og overskuelige tabeller vises de tal, som udgør grundlaget for en empirisk undersøgelse eller en rapport.
Overskuelighed er nøgleordet for en tabel, hvor uorganiserede mængder af observationer optælles og præsenteres. Tabellens grundstruktur ser ud som følger:

	Hovedspalte			
Forspalte				

Forspalten består som regel kun af én søjle, mens hovedspalten oftest består af flere (kolonner). Lad os prøve som eksempel at opstille en tabel for personer indenfor bestemte intervaller af indkomster for indkomståret 2010.

	Indkomstansættelser (1.000 DKK), personer			
	0 – 99	100 – 199	200 – 399	400 og mere
2010				

Desuden skal tabellen indholde en overskrift, som er kort, men alligevel dækkende for indholdet og måske tabelnummer. Under tabellen anføres kilden samt eventuelle noter og anmærkninger. Vi kan desuden overveje, om tabellen er god i forhold til den problemstilling, som vi ønsker at analysere. Måske vil det være mere hensigtsmæssigt at gengive den procentvise fordeling af materialet, ganske som vi gjorde i noterne om deskriptiv statistik, hvor vi også analyserede fordelingen af indkomster.

Lad os nu udvide tabellen. Vi ønsker, at at se på fordelingen af indkomsterne opgjort efter køn. Dette kan indarbejdes i tabellen, da nu vil se ud som følger:

2010	Indkomstansættelser (1.000 DKK), personer			
	0 – 99	100 − 199	200 − 399	400 og mere
Kvinder Mænd				

Hvis 2 elementer kombineres i den samme tabel, som det er gjort ovenfor, så er der tale om en *krydstabel* eller en *antalstabel.*
Når vi opstiller korrelationsdiagrammer anvender vi også materiale fra 2 fordelinger af data. *Krydstabellen* viser således også, om der er sammenhæng mellem 2 variabler. I *statistik II* vil I lære, hvordan man laver et specifikt test for at undersøge en sådan sammenhæng.
Lad os illustrere problemstillingen. Ved en spørgeskemaundersøgelse er 77 studerende ved SDU Campus Sønderborg blevet spurgt om deres foretrukne type af coca-cola. Der kunne svares mellem almindelig coca-cola og diæt coca-cola. Ved undersøgelsen angav studenterne tillige deres køn.
Svarende blev optalt og samlet i nedenstående *krydstabel,* idet totalerne også er angivet:

	Cola	Diæt cola	**Total**
Kvinde	3	33	36
Mand	30	11	41
Total	33	44	77

Ser man nærmere på tabellen, så ses der en sammenhæng mellem køn og valg af cola. Af kvinderne foretrækker 33 ud af 36 diæt colaen. Det svarende til mere end 90 %. For de mandlige studenter gælder, at 30 ud af 41 foretrækker den almindelige cola. Det svarende godt 73 %. I dette tilfælde siger man, at der er *afhængighed* mellem valg af cola og køn. Kvinderne foretrækker diæt colaen, mens mændene foretrækker den almindelig cola.
Hvordan ville tabellen se ud, hvis der ikke var sammenhæng mellem køn og valg af cola? Så ville omkring halvdelen af både kvinderne og mændene fortrække den ene af typerne. Det vil sige, at begge køn er indifferente mellem valget af de to typer af cola. Lad os prøve at rokere lidt om i tallene, så dette er gældende. Herved fremkommer følgende tabel:

	Cola	Diæt cola	**Total**
Kvinde	18	18	36
Mand	20	22	41
Total	38	39	77

I dette tilfælde siger man, at der er *uafhængighed* mellem de to variabler. Valget af typen af cola afhænger således ikke af kønnet. Måske kan andre forhold som eksempelvis alder eller indkomst tænkes at påvirke dette valg, men ikke køn.

Eksempel
Opgave 1 og 2, sommereksamen 2011.
(begge opgaver er medtaget, da materialet i opgave 1 danner udgangspunkt for opgave 2)

Opgave 1 sommereksamen 2011. (2 points, 10 %)

Tabellen viser kendetegnet antal "eksamenspoint", som 15 elever har opnået i en eksamensopgave i tysk.

Point	10	14	7	18	12	15	12	12	8	8	4	9	21	11	15

A) **1P**

Opstil den relative hyppighedstabel for kendetegnet "eksamenspoint", idet der anvendes en inddeling i målekategorier. Info: eksamensopgaven var bestået med mindst 10 point

Optællingen kan laves i et lille skema. Her har jeg bare medtaget de points, som er blevet opnået:

	4	7	8	9	10	11	12	14	15	18	21	Sum
Optælling	|	|	||	|	|	|	|||	|	||	|	|	
Antal	1	1	2	1	1	1	3	1	2	1	1	15
Relativ hyppighed	0.07	0,07	0,14	0,07	0,07	0,07	0,21	0,07	0,14	0,07	0,07	1,00
Kumuleret hyppighed	0,07	0,14	0,28	0,35	0,42	0,49	0,70	0,77	0,91	0,98	1,00	(afrundet)

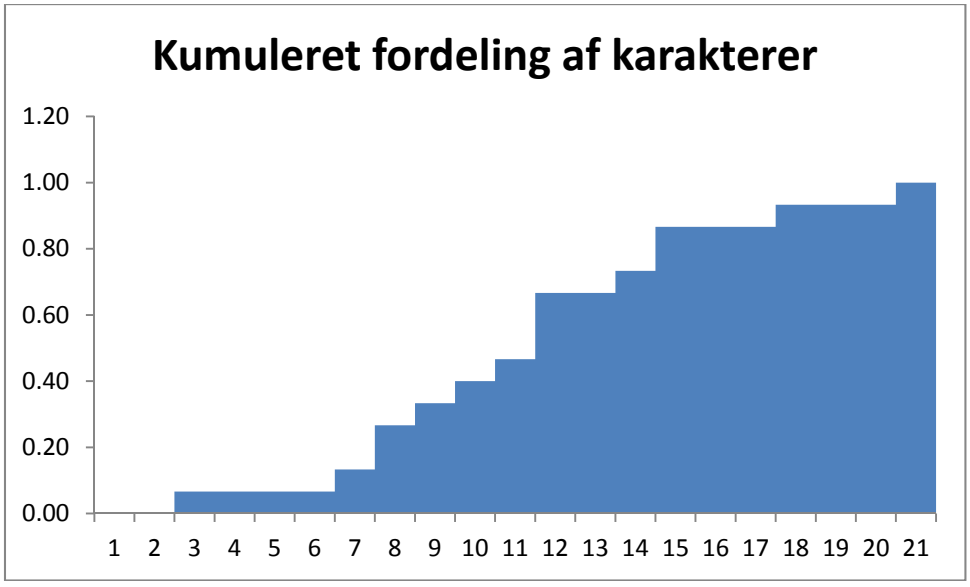
Vi kan nu gøre tabellen mindre ved kun at opdele på bestående og ikke bestående. Det ser ud som følger:

	Ikke bestået	Bestået	Total
DT	5	10	15
Relativt	0.33	0.67	1.00

B) **1P**

Tegn den kumulerede fordelingsfunktion for "eksamenspointene" så detaljeret som muligt.

Her må vi tilbage til den detaljerede fordeling. Jeg har opstillet nedenstående illustration:



Opgave 2 sommereksamen 2011. (3 points, 15 %)

I den efterfølgende tabel vises ud over eksamenspointene fra opgave 1 også "deltagelsen" (DT) i en frivillig litteratur-workshop for hver elev. Opstil et nyt kendetegn "bestået" ved at

inddele kendetegnet ”eksamenspoint“ fra opgave 1 i de 2 målekategorier ”ikke bestået“ (med point mindre end 10) og ”bestået“ (med point lig med eller større end 10).

Point	10	14	7	18	12	15	12	12	8	8	4	9	21	11	15
DT	ja	ja	ja	ja	ja	ja	nej	Nej	nej	nej	nej	nej	nej	nej	nej

A) **0,5 P**
Opstil en krydstabel for kendetegnet ”bestået“ med kendetegnet ”deltagelse“.

Jeg arbejder videre med min tabel fra opgave 1:

	4	7	8	9	10	11	12	14	15	18	21	Sum
Antal	1	1	2	1	1	1	3	1	2	1	1	20
DT		ja			ja		Ja	ja	ja	ja		6

Nu kan krydset sammentælles:

	Ikke bestået	Bestået	Total
DT	1	5	6
Ikke DT	4	5	9
Total	5	10	15

B) **1 P**
Udarbejd en velegnet procentvis beregning, og fortolk derefter tabellen mht., om der er en sammenhæng mellem at deltage i workshoppen og bestå eksamen og i givet fald hvilken.

Der er flere muligheder. Jeg har gjort som følger:

	Ikke bestået	Bestået	Total
DT	16.7 %	83.3 %	100 %
Ikke DT	44.4 %	55.6 %	100 %
Total	33.3 %	66.7 %	100 %

Af de studerende med DT var beståelsesprocenten godt 83, mens den for studenterne uden var på godt 55 %. Fordelingen mellem bestående og ikke bestående af studenter, der ikke har DT er imidlertid meget ens.

C) **1 P**
Hvilke værdier skulle stå inde i krydstabellen, hvis den skulle vise, at der ikke er nogen sammenhæng mellem deltagelse i workshoppen og chancen for at bestå eksamen?

	Ikke bestået	Bestået	Total
DT	3	3	6
Ikke DT	4	5	9
Total	7	8	15

Så skulle fordelingen mellem ikke bestået og bestået med DT være mere identiske. Eksempelvis som ovenfor.

D) **0,5 P**
Kan man ud fra sammenhængen mellem begge variabler generelt konkludere, at det ville være et godt tiltag at gøre deltagelsen i workshoppen obligatorisk for at øge chancen for at flere kan bestå eksamen?

Det er vanskeligt at sige helt tydeligt, da materialet er meget lille, men meget tyder på en positiv effekt.

Nils Karl Sørensen ©
Statistik II, BA int, Flensborg

Sæt 1: Opstilling af hypoteser og udførsel af simple tests

af Nils Karl Sørensen

Indhold *side*

1. Formulering af hypoteser og fejlkilder

Afprøvning af hypoteser er nok den mest hyppige from for analyse i statistik.

- Hypoteseafprøvning kan anvendes til at undersøge validiteten af en given *påstand* i forhold til et statistisk materiale, som vi har indsamlet
- Hypoteseafprøvning kan forekomme for såvel middelværdien som for variansen. I nærværende sæt af noter ses der alene på middelværdien og populationsandelen, mens der i det følgende sæt af noter også ses på variansen
- Hypoteseafprøvning kan finde sted enten i totalpopulationer eller i stikprøver. I førstnævnte siger man at standardafvigelsen σ er kendt, mens man i stikprøver siger at σ ikke er kendt. I stedet kender man stikprøvens standardafvigelse betegnet s

I de simpleste tests med hypoteser sammenlignes et givet datasæt i forhold til en givet *påstand.* I det følgende sæt af noter udvides problemstillingen til at omfatte en sammenligning af 2 datasæt eller stikprøver. Endelig er variansanalysen, som behandles i det tredje sæt noter, en sammenligning af p datasæt eller stikprøver.

En *hypotese* er en antagelse, som i virkeligheden er sand eller falsk (kan accepteres eller forkastes). Hypotesetestet går således ud på, at vælge mellem 2 hypoteser benævnt H_0 og H_1 (også kaldet H_a, hvor "a" står for "alternativ"). Definer hypoteserne som:

- **Nul hypotesen H_0** som betegner basisscenariet eller *status quo.* Basisscenariet er den undersøgelse, som man allerede har indsamlet materiale for. Den er sand indtil noget andet er bevist
- **Alternativ hypotesen H_1** som er den *påstand* man søger at undersøge validiteten af

Hovedpunkterne i et hypotese test er:

1. Identifikation af problem og formulering af de to hypoteser
2. Valg af en relevant procedure for testet (dette kaldes også *testeren*)
3. Beregning af testeren
4. Fastlæggelse af *signifikansniveauet* α (hvis ikke noget er angiver er det altid $\alpha = 0.05$)
5. Evaluering af udfaldet af testeren i forhold til en relevant statistisk fordeling
6. Accept eller forkastelse af nul hypotesen (H_0) eller alternativ hypotesen (H_1)

Man kan sige, at *signifikansniveauet* udtrykker den grad af *forskellighed*, som er acceptabel i den givne situation. Antager man eksempelvis et signifikansniveau på 95 procent, så betyder det, at hvis man accepterer H_1 hypotesen, så er *påstanden* så forskellig fra basisscenariet, at dette kun vil ske i 5 procent af alle tilfælde. *Påstanden* er så ekstrem, at den kun vil være lig med værdien i basisscenariet i 5 tilfælde ud af 100.

Dette ligner meget det, som vi arbejdede med, i sættet af noter omhandlende *konfidensintervaller*. Hvis H_1 blev accepteret, så faldt *påstanden* udenfor konfidensintervallet, fordi den var for afvigende eller ekstrem i forhold til basisscenariet.

Hidtil har opstilling af hypoteser begrænset sig til at omfatte nulhypotesen og alternativhypotesen. Imidlertid er der være en række overvejelser, som man skal gøre sig, når man opstiller hypoteser. Lad os se på disse.

Diskussion af hypoteser og fejlkilder

For at kunne se på de metodiske aspekter af hypoteseafprøvning, så må vi en tur i retten! I retten forelægges og afprøves hypoteser om skyld og uskyld for dommer og nævninge. Disse når via loven og retspraksis frem til domfældelsen.

Som udgangspunkt er den tiltalte altid *uskyldig*. Dette er basisscenariet H_0. Forsvareren skal forsvare basisscenariet, mens anklageren skal argumentere for *alternativet* H_1. Dommeren fastlægger via loven og retspraksis *signifikansniveauet.* Er den forseelse, som man er tiltalt for, signifikant, vil den tiltalte blive dømt skyldig, og der vil blive udmålt en straf.

Det lyder jo pænt og ordentligt, men erfaringen viser, at retten ikke er uden fejl. Disse fejl findes også i hypotesetestene. Lad os se på disse.

Vi betragter en fodgænger, som kl. 02:00 fredag krydser Aabenraade Strasse i Flensborg for rødt. Det bliver observeret af trafikpatruljen, og fodgængeren bliver noteret og bringes i retten. Som udgangspunkt er fodgængeren uskyldig, indtil det modsatte er bevist. Hvis dette udsagn passer, så er hypotesen *H_0 sand,* og den kan ikke forkastes. Med dette udgangspunkt kan der opstilles følgende hypoteser for færdselsforseelsen:

H_0: Ikke skyldig i at gå over for rødt lys
H_1: Skyldig i at gå over for rødt lys

Imidlertid kan det også være, at fodgængeren aktuelt krydsede vejen, mens der var rødt lys. Dette blev observeret af trafikpatruljen, og vidneudsagnet indikerer, at personen er skyldig. Dommeren følger vidnet og kender fodgængeren skyldig. I denne situation forkastes H_0, mens H_1 accepteres og er sand.

I disse to tilfælde tages således den korrekte beslutning. Der er overensstemmelse mellem hypotesen omkring trafikforseelsen, og den afsagte kendelse. Disse observationer er gengivet i tabellen nedenfor.

	Sagens rette tilstand	
	H_0 er sand	H_1 er sand
H_0 accepteres	Korrekt beslutning	Type II fejl (β)
H_1 accepteres	Type I fejl (α)	Korrekt beslutning

Som det fremgår, er der imidlertid to andre muligheder. Disse muligheder benævnes henholdsvis en fejl af type I (α) og en fejl af type II (β).

Betragt først **type I fejlen**. Den kaldes også for α-fejlen. Her er sagens rette tilstand, at H_1 accepteres, mens at *H_0 er sand*. Fodgængeren findes således skyldig i, at have gået over for rødt lys, men det var en anden som begik forseelsen. Der er tale om et *justitsmord.* Dette kunne være tilfældet, hvis vidnet ikke havde observeret godt nok, eller forsvaret var ringe. Dommeren blev ikke overbevist, og afsagde den forkerte kendelse.

Er justitsmordet det værste tilfælde? Betragt en anden og måske mere alvorlig fejl i det følgende og døm selv!

Type II fejlen kaldes også for β-fejlen. Her tiltales en person for trafikforseelsen og idømmes straf. Personen er imidlertid en helt anden person uden tilknytning til hændelsen. Den virkelige skyldner bliver aldrig tiltalt og går fri! For fodgængeren gælder således, at H_1 er sand (skyldig), men H_0 accepteres (personen anholdes ikke og kommer aldrig i retten).

Lad os prøve at anskue dette lidt anderledes. Træneren til en håndboldkamp i Campushalle har en forventning om, at modstandernes målvogter er dårlig. Træneren forklarer spillerne, at det er let at lave mål. Det viser sig imidlertid, at målvogteren er bedre end sit rygte, og tager rigtig mange skud, så kampen tabes. Hvad betyder dette i en statistisk kontekst? Træneren anvender en forkert model til at forudsige kampens udfald. Den middelværdi, som træneren anvender, til at forudsige målvogterens redningsprocent, er systematisk undervurderet. Modellen er således misvisende og vil give systematisk forkerte resultater.

En *styrkefunktion* angiver sandsynligheden for ikke at begå en type II eller β-fejl i forhold til den sande model. Modellering af styrkefunktionen er ikke en del af pensum.

Tabellen nedenfor opsummerer diskussionen om fejl og hypoteser:

	Sagens rette tilstand	
	Ikke skyldig	Skyldig
Ikke skyldig	Korrekt beslutning	Type II fejl (β)
Skyldig	Type I fejl (α) Justitsmord	Korrekt beslutning

De to typer af fejl α og β skal minimeres. Opskrevet som betingende sandsynligheder haves at:

α = P(type I fejl) = P(forkast H_0 | H_0 er korrekt)

Man forkaster en nulhypotese selv om denne er sand

β = P(type II fejl) = P(accepter H_0 | H_0 er falsk)

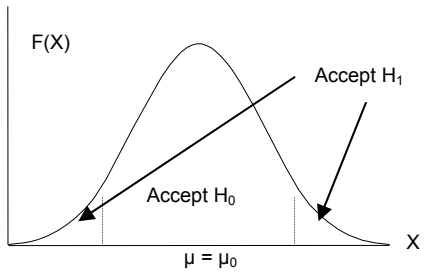
Man accepterer en nulhypotese, som er falsk

Opstilling af hypoteser

Betragt som et eksempel et test for, at middelværdien μ_0 i et givet datasæt er lig med en postuleret værdi. For et sådant test, kan der opstilles hypoteser på tre måder; nemlig som tosidet og ensidet herunder enten som venstresidet eller som højresidet.

1) Tosidet test H_0: $\mu = \mu_0$ (middelværdien er lig med basisværdien)
H_1: $\mu \neq \mu_0$ (middelværdien er *forskellig* fra basisværdien)

Illustration:

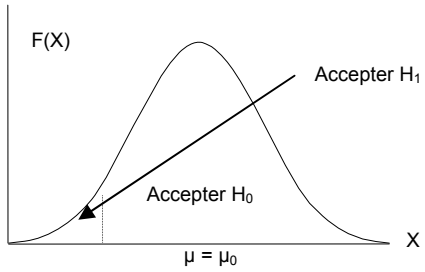


Lad os nu vende os mod de to ensidede tests

2) Venstresidet test

H_0: $\mu \geq \mu_0$ (middelværdien er identisk eller større end basisværdien)
H_1: $\mu < \mu_0$ (middelværdien er mindre end basisværdien)

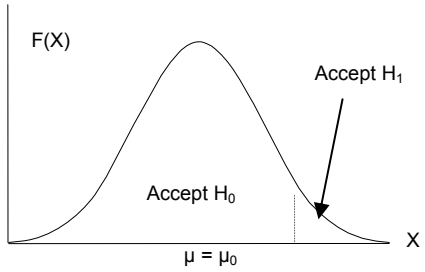
Illustration



3) Højresidet test

H_0: $\mu \leq \mu_0$ (middelværdien er identisk eller mindre end basisværdien)
H_1: $\mu > \mu_0$ (middelværdien er større end basisværdien)

Illustration



Det observeres, at test af hypoteser har meget tilfælles med opstilling af konfidensintervaller.

- I det *tosidede tilfælde* er sandsynligheden for at acceptere H_0 lig med $(1-\alpha)$, mens sandsynligheden for at acceptere H_1 er lig med $\alpha/2$ (det er den øvre og nedre grænse i konfidensintervallet)
- I det *ensidede* tilfælde er sandsynligheden for at acceptere H_0 lig med $(1-\alpha)$, mens sandsynligheden for at acceptere H_1 er lig med α

2. Z - baserede og T - baserede tests for middelværdien

Z - baseret test med σ kendt

Betragt først et test baseret på normalfordelingen. Det vil sige på *Z*. Det antages, at der er tale om en totalpopulation, hvor standardafvigelsen σ er kendt. Der kan nu opstilles følgende hypoteser og tester for eksempel i det tosidede tilfælde:

Hypoteser: $H_0: \mu = \mu_0$
$H_1: \mu \neq \mu_0$

Under nul hypotesen er modellen korrekt, og identisk med den værdi, der testes for, mens det under den alternative hypotese antages, at den værdi, der testes for, ikke er konsistent med datasættet.

Testeren bliver nu:

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

Her er $\overline{X}$ den værdi, der testes for, mens μ_0 er middelværdien i datasættet under H_0. Endvidere betegner σ standardafvigelsen, og n er antallet af observationer i datasættet. Testeren fremkommer som en omformulering af den centrale grænseværdisætning kendt fra noterne til Statistik I.

Eksempel: Prisen på benzin

Vi har i noterne til Statistik I beskæftiget os med et eksempel omhandlende prisen på benzin, som blev antaget at følge en normalfordeling. Baseret på 25 observationer fandtes en middelværdi på 9.95 DKK med en standardafvigelse på 0.30 DKK. Vi undersøger nu følgende påstand:

- Prisen på benzin er på en tilfældig dag lig med 10.10 DKK

Det materiale på 25 observationer, som er indsamlet ved besøg på benzinstationer, antages at være status quo eller basismaterialet. Det vil sige, at det er gældende under H_0. Hypoteserne kan da opstilles som:

$H_0: \mu = 10.10$ (prisen er lig med 10.10 DKK)
$H_1: \mu \neq 10.10$ (prisen er *forskellig* fra 10.10 DKK)

Vi har ikke sagt noget specifikt i udgangssituation, så der antages et tosidet test med et signifikansniveau på 95 procent. Det vil sige, at det antages at $\alpha = 0.05$. Da testet er tosidet haves at $\alpha/2 = 0.025$.

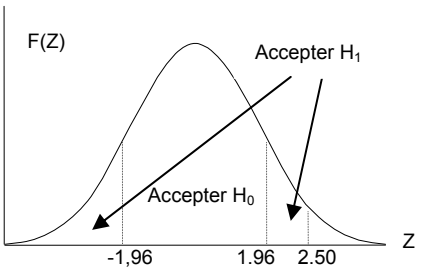
Testeren findes idet $\overline{X} = 10.10$ DKK ved indsættelse af værdierne i udtrykket ovenfor som:

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}} = \frac{10.10 - 9.95}{0.30 / \sqrt{25}} = \frac{0.15}{0.06} = 2.50$$

De kritiske værdier findes i Z-fordelingen i **Statistics Tables.** Ved et signifikansniveau på 5 procent er de, ganske som det blev fundet, da vi arbejdede med konfidensintervaller. Det vil sige $Z = \pm 1.96$.

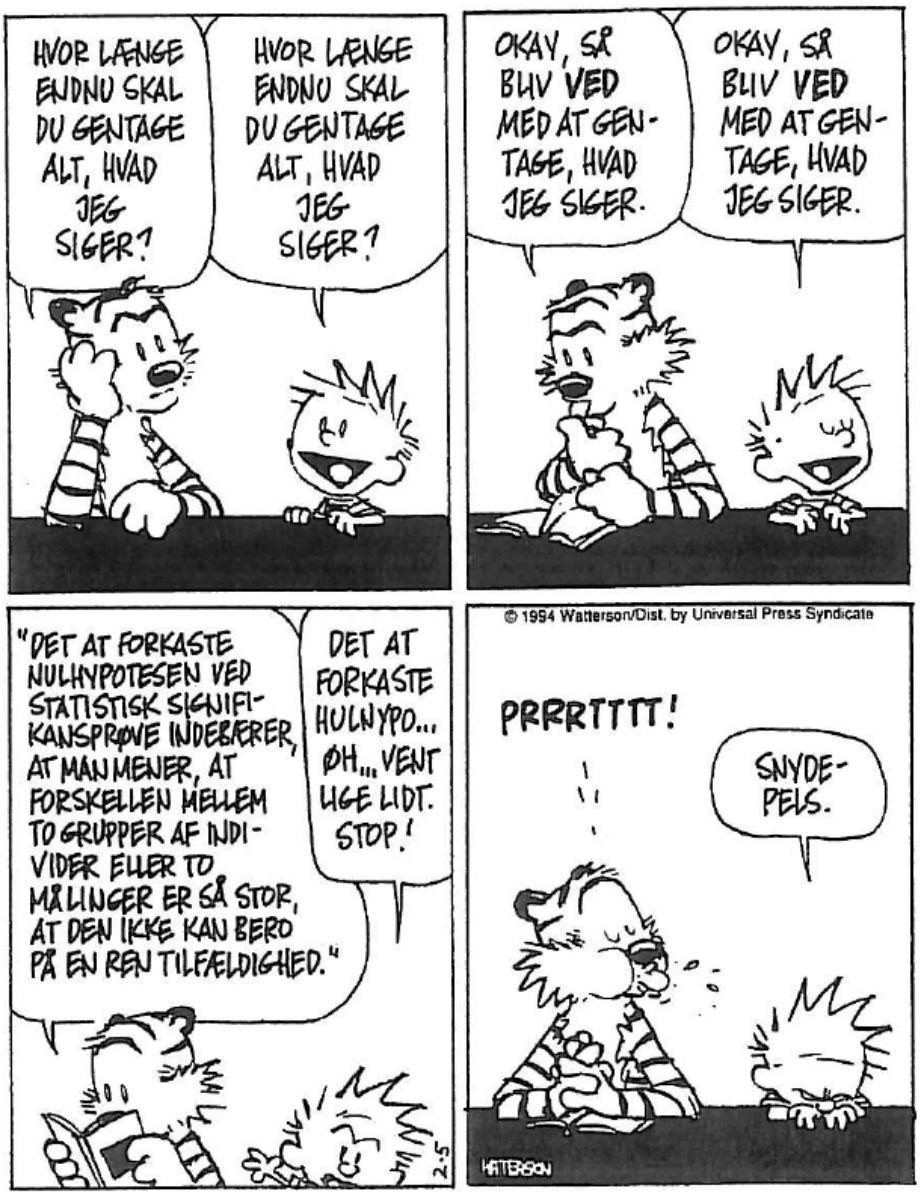
Da $2.50 > 1.96$ *forkastes* H_0. Det betyder, at prisen på 10.10 DKK ved et 5 procent signifikans er forskellig fra den gennemsnitlige pris i datasættet på 9.95 DKK. Prisen er således ikke acceptabel for bilisten, der kun vil tanke op, hvis der næsten ikke er benzin på bilen til at fortsætte rejsen.

Eksempelet kan illustreres som følger:



Bemærk, at hvis vi havde vendt testet om, så have vi fået at $Z = -2.50$. Da $-2.50 > -1.96$ vil man få den samme konklusion!

Signifikans betyder forskelligt – se øverst på næste side ☺.



P-værdien

Hvad vil det sige, at en værdi er signifikant? Vi kan ved anvendelse af tabellen for normalfordelingen beregne, *hvornår* afstanden mellem 10.10 DKK og 9.95 DKK er så stor, at afvigelsen bliver signifikant.

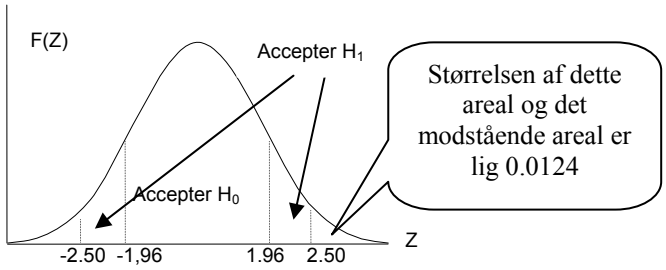
Ovenfor fandt vi, at afvigelsen bliver signifikant, netop når Z passerer 2.50 eller −2.50. Vi skal have begge sandsynligheder med, da testet var udført som et tosidet test. Ved hvilken sandsynlighed indtræffer dette? Der fremkommer følgende:

$$P(Z < -2.50) + P(Z > 2.50) = 2P(Z < -2.50) = 2(0.0062) = 0.0124 \quad (=p)$$

Sandsynligheden på 0.0062 kommer ved opslag i **Statistics Tables** på Z = −2.50. Da den positive værdi på Z = 2.50 også skal medtages ganget med 2. Dette kan vi gøre, da normalfordelingen er symmetrisk.

P-værdien er lig med 0.0124. Hvad betyder dette? Ved en sandsynlighed på mindre 1.24 procent vil man forkaste H_0, selvom denne er sand. Denne sandsynlighed betegner således risikoen for at begå en *type I fejl*.

Man kan illustrere dette som følger:



P-værdien anvendes til at sammenligne med det signifikansniveau, der er på forhånd er valgt for undersøgelsen. Dette er typisk lig med 95 procent svarende til at $\alpha = 0.05$. Hvis p-værdien er mindre end værdien for α, så forkastes H_0. Typisk kan man indlægge flere signifikansniveauer i sin undersøgelse.

Dette er helt centralt, og anvendes meget i såvel tests som i regressionsanalyse, der er emnet for notesæt 5. Rammen nedenfor opsummerer, hvad p-værdien er, og viser nogle signifikansniveauer.

P-værdien:

- Hvis p-værdien er mindre end det signifikansniveau, som man opererer med, må nulhypotesen forkastes

- Desto mindre en p-værdi man beregner, desto mindre er sandsynligheden for, at nulhypotesen er sand

Hvis den beregnede p-værdi under nulhypotesen er *mindre* end

- For **p < 0.10** haves ***svag signifikans*** for at H_0 forkastes
- For **p < 0.05** haves ***signifikans*** for at H_0 forkastes
- For **p < 0.01** haves ***stærk signifikans*** for at H_0 forkastes

Opgives der ikke noget i en opgave, så antages det altid at $\alpha = 0.05$, og der opereres med signifikans.

I eksemplet ovenfor om benzinprisen fandtes, at p-værdien var lig med 0.0124. Da den er mindre end 0.05 er afvigelsen signifikant, men da $0.0124 > 0.01$ fandtes der ikke stærk signifikans.

T - baseret test med σ ikke kendt

På tilsvarende vis kan man opstille et test, hvor populationens standardafvigelse ikke er kendt. Dette vil være tilfældet i en strikprøve eller i et mindre datasæt. I dette tilfælde benævnes standardafvigelsen i stikprøven *s*. Testeren kan opskrives på tilsvarende måde for et tosidet test, som ved Z-testet.

Hypoteser: $H_0: \mu = \mu_0$
$H_1: \mu \neq \mu_0$

Under nul hypotesen er modellen korrekt, og identisk med den værdi, der testes for, mens det under den alternative hypotese antages, at den værdi, der testes for, ikke er konsistent med datasættet.

Testeren bliver nu:

$$t = \frac{\overline{X} - \mu_0}{s / \sqrt{n}}$$

Her er $\overline{X}$ den værdi, der testes for, mens μ_0 er middelværdien i datasættet under H_0. Endvidere betegner *s* standardafvigelsen i stikprøven eller det mindre datasæt, og *n* er antallet af observationer i datasættet.

Hvad er et mindre datasæt? I de fleste lærebøger i statistik angives det til at være mellem 25 og 30 observationer. Det er dog en smagssag. Fremkomsten af PC'ere med rigelig med regnekapacitet har betydet, at man i dag siger, at "mindre" er op til 120 observationer. For nærværende anvender vi dog det førstnævnte og bliver på en 25 til 30 observationer.

Testeren hedder nu "t" og den er ***student t-fordelt*** (eller blot t-fordelt) med *frihedsgrader* lig med *fg = (n−1)*. T-fordelingen er en variant af normalfordelingen, der er specielt velegnet til små datasæt. T-fordelingen er tabuleret i **Statistics Tables** på side 10. Et ekstrakt af fordelingen, som den er gengivet i **Statistics Tables** findes øverst på næste side.

Hvad betyder frihedsgrader? Det er det samme som muligheder/observationer fratrukket begrænsninger. I nærværende situation er begrænsningen, at vi befinder os i en stikprøve eller et mindre datasæt. Her har vi beregnet en middelværdi, og det koster en frihedsgrad. Derfor *(n−1)*. Hvis vi eksempelvis arbejdede med 3 stikprøver, ville frihedsgraderne være lig med *n* minus 3.

	t Values				
df	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
…	…	…	…	…	…
…	…	…	…	…	…
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
…			…	…	…
…	…	…	…	…	…
∞	1.282	1.645	**1.96**	2.326	2.576

Bemærk den nederste linje med det uendelige antal frihedsgrader. Her finder vi z-værdierne, som de kendes fra normalfordelingen. Det betyder, at *t-fordelingen tilnærmes normalfordelingen, når antallet af frihedsgrader øges.* Dette understreger, at denne fordeling er en transformation af normalfordelingen.

Bemærk også at det kun er et lille udsnit af fordelingen, som er med i tabellen. Faktisk er der for hver frihedsgrad en hel bagvedliggende normalfordeling. Da vi imidlertid kun anvender en lille del af tabellen, så er en udformning af tabellen som ovenfor, det mest hensigtsmæssige.

Da t-fordelingen er en transformeret normalfordeling har fordelingen middelværdien nul, mens variansen for antal frihedsgrader større end 2 er lig med $\sigma = fg/(fg-2)$. Bemærk at når *fg* øges i det uendelige, så går variansen mod én ganske som under normalfordelingen.

Hvorfor kaldes det en *Student t fordeling?* I 1908 opdagede forskeren og statistikeren W.S. Gossett, som var ansat ved Guinness bryggeriet i Dublin, denne fordeling. Han var blandt andet beskæftiget med kvalitetskontrol. Han havde fundet ud af, at virksomhedens kvalitetskontrol af produkterne systematisk var behæftet med fejl i de mindre stikprøver, der blev udtaget til kontrol. Han fandt, at man for ofte kasserede varer i forhold til det, som man forventede i forhold til normalfordelingen. For at løse dette problem udviklede han en variant af normalfordelingen. Bryggeriet ville imidlertid ikke tillade, at han publicerede sit forskningsresultat i eget navn. Derfor anvendte han synonymet *Student t.* Denne fordeling er, som det vil ses senere i kurset, ekstremt anvendelig og meget udbredt!

Lad os nu beregne testeren i eksempelet med benzinprisen. Testeren er ganske som under Z-testet. Vi antager her, at vi har at gøre med en stikprøve fra en uendelig proces af besøg/påfyldninger på en tankstation. Testen bliver nu:

$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} = \frac{10.10 - 9.95}{0.30/\sqrt{25}} = \frac{0.15}{0.06} = 2.50$$ (som ovenfor)!

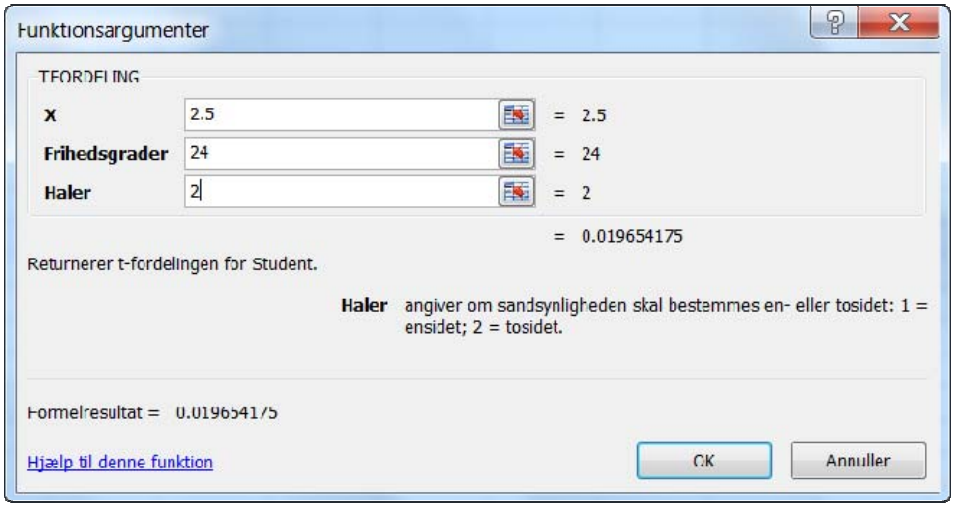
Imidlertid er testeren nu *t-fordelt med antallet af frihedsgrader lig med fg = (n−1) = (25−1) = 24.* Antag et signifikansniveau på 95 procent svarende til α = 0.05, og med et tosidet test svarende til α/2 = 0.025 findes ved anvendelse af **Statistics Tables** at t-værdien er lig med *t = 2.064.* I tabeludtaget for t-fordelingen på den foregående side er dette vist, og den kritiske værdi er markeret med gult.

Da *2.064 < 2.50* forkastes nulhypotesen, og vi får den samme konklusion som ovenfor.

Prøv at se på tallene! Vi finder, at t-værdien er lig med 2.064 for datasættet på 25 observationer. Hvis vi nu havde et datasæt, som var uendelig stort, så ville vi få en t-værdi (og z-værdi) på 1.96. Ved anvendelse af t-fordelingen bliver spredningen en anelse større. Anvendelse af t-fordelingen i små stikprøver reducerer således sandsynligheden for en type I fejl i forhold til en situation, hvor normalfordelingen var blevet anvendt.

Kan man finde p-værdien for testet, hvis det er t-fordelt? Det er muligt, men da der er en hel normalfordeling bag ved hver frihedsgrad i t-fordelingen, så er det ikke så ligetil. Fra tabellen ovenfor fremgår det, at ved 24 frihedsgrader er $t_{\alpha=0.01} = 2.492$. Da *2.492 < 2.50,* har vi også stærk signifikans (i modsætning til ved anvendelse af normalfordelingen).

Vi kan imidlertid anvende Excel. Brug ***formler/indsæt funktion/statisik/tfordeling*** og nedenstående menu fremkommer. Indsæt 2.5 for X. Husk 2 = tosidet og 1 = ensidet. Så findes det at *p= 0.019* jævnfør nedenstående skærmbillede.



3. Z - baseret test for populationsandelen

I tilfældet med en totalpopulation eller et stort datasæt kan der opstilles et test for populationsandelen efter samme retningslinjer som for middelværdien. Dette test er en udvidelse af det i Statistik I gennemgåede konfidensinterval for populationsandelen p defineret som $p = x/n$, hvor x er en variabel med et givet karakteristika eksempelvis at stemme "ja" ved en afstemning, mens n er antallet af observationer i datasættet.

Der kan nu opstilles et test for eksempelvis, at andelen er lig med en given værdi.

Hypoteserne er: $H_0: p = p_0$
$H_1: p \neq p_0$

Der kan nu opstilles følgende tester:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Som tidligere angiver Z normalfordelingen, mens p_0 er andelen under H_0 dvs. i udgangssituationen. $\hat{p}$ angiver den værdi eller påstand, som der testes for. Ganske som det var tilfældet ved opstillingen af konfidensintervallet ses det, at testet bygger på en betragtning fra binomialfordelingen med to udfald eksempelvis at stemme "ja" eller "nej". For at testet skal gælde forudsættes at np_0, $n(1-p_0)$ begge er større end 5.

Eksempel: Bageren på Langeland
Tilbage i Statistik I beskæftigede vi os med bageren på Langeland. Han lavede kager, hvor man i glasur kunne få "ja" eller "nej" afhængigt af, hvad man ville stemme omkring sammenlægningen af de kommuner på øen til én enkelt kommune..

Optil valget solgte han i alt 656 kager. Af disse bar 385 "ja", mens 271 blev dekoreret med et "nej".

Antag nu, at en borgmester i en af kommunerne er af den opfattelse, at der vil være 63 procent af de afgivne stemmer som er "ja" til sammenlægningen.

- Anvend nu de oplysninger, der er fremkommet om det antal kager, som bageren har solgt til ved et 95 procents signifikansniveau at undersøge validiteten af borgmesterens påstand

Først opstilles hypoteserne: $H_0: p = 0.63$ (påstanden er korrekt)
$H_1: p \neq 0.63$ (påstanden er *ikke* korrekt)

Nu opstilles testeren i det $p_0 = x/n = 385/656 = 0.587$ mens $\hat{p} = 0.63$:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.63 - 0.587}{\sqrt{\frac{0.587(1-0.587)}{656}}} = \frac{0.043}{0.0192} = 2.24$$

Vi har opstillet et tosidet test, og nu skal vi finde den kritiske værdi ved anvendelse af tabellen for normalfordelingen.

Der anvendes et 95 procents signifikansniveau. Det vil sige at $\alpha = 0.05$ og $(1-\alpha) = 0.95$. Hvis testet er tosidet, så er $\alpha/2 = 0.025$. I tabellen for normalfordelingen i **Statistics Tables** findes værdien for Z til ±1.96. Da 2.24 > 1.96 forkastes H_0. Det vil sige, at 0.63 er signifikant forskellig fra 0.587.

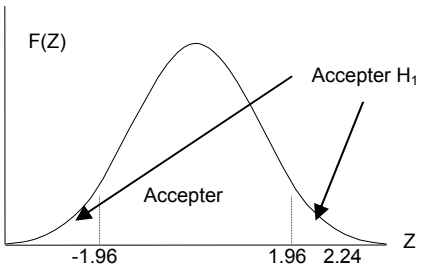
Borgmesterens påstand er således ikke korrekt.

Hvad er *p-værdien?* Vi kan finde denne ved anvendelse af normalfordelingstabellen i **Statistics Tables**. Her fås:

$$P(Z < -2.24) + P(Z > 2.24) = 2P(Z < -2.24) = 2(0.0125) = 0.0250 \ \textit{(=p)}$$

Den kritiske værdi er således på 0.025 svarende til 2.5 procent. Da $p < 0.05$ er testet signifikant ved 5 procent og H_1 *accepteres.*

Illustration:



Hvad nu hvis vi havde haft et signifikansniveau på én procent? I dette tilfælde er 0.025 > 0.01. Dvs. at $p > 0.01$ og konklusionen er omvendt dvs. at H_0 accepteres og borgmesterens påstand er korrekt. Dette eksempel viser, at valget af signifikansniveau har betydning for udfaldet af testet. Dette er en væsentlig konklusion.

Bemærk også, at resultatet af testet kunne være opnået ganske, som vi gjorde i Statistik I. Her så vi på påstanden i forhold til konfidensintervallet. Falder værdien udenfor intervallet forkastes H_0, mens H_0 accepteres, hvis værdien falder indenfor konfidensintervallet.

I noterne om opstilling af konfidensintervaller fandt vi 95 % konfidensintervallet til at blive lig med $[0.549\,;0.625]$. Det vil (som forventet) sige at 0.63 falder udenfor intervallet og påstanden er falsk.

Betragt til sidst et test, der en ensidet. I dette tilfælde er $\alpha = 0.05$ og den tilsvarende Z-værdi er lig med ± 1.645. Da $1.645 < 2.24$ ændres der ikke ved konklusionen.

Får man altid den samme konklusion ved at gå fra et tosidet til et ensidet test? Nej, betragt en situation, hvor udfaldet af testeren Z_{test} ligger mellem $1.645 < Z_{test} < 1.96$. I et sådant tilfælde accepteres H_1 ved et ensidet test, mens H_0 accepteres i et tosidet test. Dette understreger vigtigheden af, at formulere testet korrekt.

Nils Karl Sørensen ©
Statistik II, BA int, Flensborg

Sæt 2: Hypoteser og tests i to uafhængige stikprøver

af Nils Karl Sørensen

Indhold *side*

1. Sammenligning af to middeltal

Nu udvides analysen til at omfatte en sammenligning af to uafhængige stikprøver, datasæt eller populationsandele. Sammenligningen kan være for middelværdien, andelen eller variansen. De 2 datasæt er nummereret med fodtegn *1* og *2*. Det antages, at begge datasæt er normalfordelte. Oftest er de to datasæt af forskellig størrelse med antal observationer lig med n_1 og n_2. Det vil sige, at der som regel gælder at $n_1 \neq n_2$.

Eksempler på sammenligninger af to datasæt finder ofte anvendelse indenfor marketing, produktionskontrol eller medicin. Eksempler herpå kan være:

- Undersøgelse af effekten af en markedskampagne, hvor markedsandelen på et givet produkt sammenlignes før og efter kampagnen
- Undersøgelse af effekten af indførslen af en ny produktionsproces eller lignende
- Undersøgelse af effekten af en ny type medicin mellem en test- og en referencegruppe
- Undersøgelse af om volatiliteten på to typer af aktiver er den samme, ved at sammenligne de to varianserne for de to aktiver

Som i det foregående sæt af noter ses der på to forskellige tilfælde, når det gælder tests for identiske middelværdier; dels testet i totalpopulationer, som er et Z-test, dels testet i stikprøver, som er et t-test. I førstnævnte test er standardafvigelsen σ kendt, mens dette ikke er tilfældet idet sidstnævnte test, hvor σ erstattes af stikprøvens standardafvigelse kaldet *s*.

Testet for sammenligninger af andele er et Z-test. Det er en udvidelse af det tilsvarende test for en enkelt andel fra det foregående sæt af noter. Endelig er testet for sammenligning varianserne lidt mere specielt, da der sammenlignes for to kvadrerede variabler σ^2.

Z - baseret test med σ kendt

Man betragter to store datasæt eller totalpopulationer, således at normalfordelingen vil være gældende. For at lette gennemgangen antages testet at være tosidet. Hypoteserne for identiske middelværdier er, idet datasættene kaldes henholdsvis 1 og 2:

Hypoteser:

H_0: $\mu_1 = \mu_2$	(Middeltallet i de 2 datasæt er ens)
H_1: $\mu_1 \neq \mu_2$	(Middeltallet i de 2 datasæt er forskellige)

eller

H_0: $\mu_1 - \mu_2 = D_0$	(Forskellen i middeltallene er lig differencen)
H_1: $\mu_1 - \mu_2 \neq D_0$	(Forskellen i middeltallene er ikke lig differencen)

Hvor D_0 er en hypotetisk forskel eller difference. Normalt vil denne antage værdien nul. Hvad kan man anvende dette til? D_0 siger, at der er en forskel i niveauet mellem de to undersøgte datasæt. Det kunne være en lønforskel opgjort på køn. Selvom det i mange lande hævdes, at der er ligeløn, så kan virkeligheden måske være anderledes. En undersøgelse kan eksempelvis indikere en lønforskel på 1.000 € om året mellem kønnene for udførslen af det samme arbejde. Dette forhold kan indbygges i de undersøgte hypoteser som følger:

H_0: $\mu_{mand} - \mu_{kvinde} = 1.000$	(lønforskellen er 1.000 €)
H_1: $\mu_{mand} - \mu_{kvinde} \neq 1.000$	(lønforskellen er ikke 1.000 €)

Hvad er fortolkningen af disse hypoteser? Accepteres H_0, så er middelværdien for mænd 1.000 € højere end for kvinder. Den øvrige stokastiske variation er ikke signifikant. Accepteres H_1, så er forskellen mellem middelværdien for mænd større end mindre end 1.000 €. Bemærk at testet kan gøres mere specifikt ved eksempelvis at opstille et ensidet test. Hypoteserne kunne eksempelvis opstilles som:

H_0: $\mu_{mand} - \mu_{kvinde} \leq 1.000$	(lønforskellen er 1.000 € eller mindre)
H_1: $\mu_{mand} - \mu_{kvinde} > 1.000$	(lønforskellen er over 1.000 €)

Her er anvendt et højresidet test jævnfør side 5 i det første sæt af noter til Statistik II.

Det Z–baserede test for identiske middelværdier er en udvidelse af testet på side 6 i det første sæt af noter til Statistik II. Normalt antages D_0 at være lig nul. Testeren bliver nu, idet n_1 og n_2 angiver antallet af observationer i henholdsvis datasæt 1 og datasæt 2; σ_1^2 og σ_2^2 er varianserne, mens $\overline{X}_1$ og $\overline{X}_2$ er middelværdierne:

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

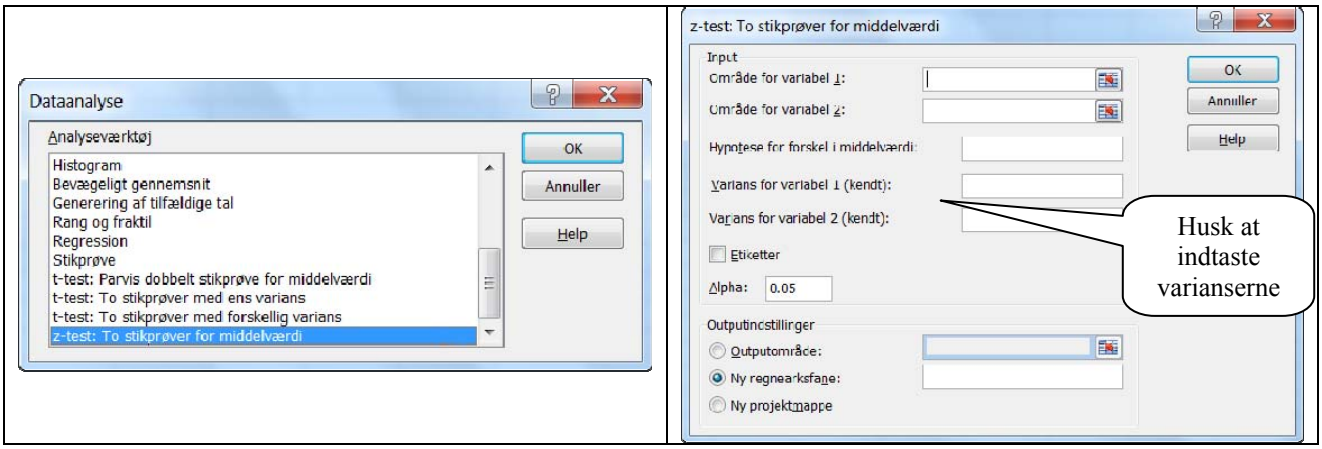
Et konfidensinterval for forskellen mellem de to middelværdier kan beregnes som:

$$\left[(\overline{X}_1 - \overline{X}_2) \pm Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]$$

For et 95 procent konfidensinterval haves at $\alpha = 0.05$. Det vil sige $\alpha/2 = 0.025$. For et 95 procents interval er $Z = \pm 1.96$.

Testeren kan beregnes enten direkte ved indsættelse eller ved anvendelse af en lommeregner Texas TI-84/89.

I Excel anvendes ***Data/Data analyse/Z-test: To Stikprøver for middelværdi,*** som vist nedenfor til venstre. Marker "OK", og så vises billedet til højre. Her skal man indtaste varianserne manuelt (testet forudsætter jo at disse er kendt på forhånd), mens data kan markeres om "område"-boksene for de 2 dataserier. Fortolkningen af udskriften vil blive behandlet senere i dette sæt noter.



På ***Texas TI-84/89*** lommeregneren er det lidt anderledes, da der er flere muligheder. På **TI 89 lommeregneren** skal man starte lidt anderledes. Når lommeregneren startes findes menuen *Stat/Lists* og der trykkes på *Enter*[3]. Nu kommer man ind i en menu magen til den, der er i **TI-84** lommeregneren. I denne menu finder man "F6 Tests" i toppen (tast f.eks. først på F1 og anvend dernæst piletasterne)

På **TI-84 lommeregneren** tastes "stat" → "Tests" → "3: 2-SampZtest" → Stats → (nu indsættes værdierne for middelværdierne og standardafvigelserne).

[3] Man kan nu få et spørgsmål om placering af data, som man svarer "OK" til og trykker "enter". Indtastning af data finder sted i registrene *L1, L2* osv. Arbejder man med frekvenser, så indtastes værdien f.eks. karakteren "7" i *L1,* mens frekvensen/hyppigheden f.eks. antallet "10" indtastes i *L2.*

Alternativt til "stats" kan man anvende "data", og anvende de data, som man eksempelvis har indtastet i registrene *L1* og *L2*.

Man kan vælge mellem etsidet og tosidet test. Endelig vælges "calculate" → Nu fremkommer den beregnede Z-værdi. Denne sammenlignes nu med den kritiske Z-værdi fra normalfordelingen i **Statistics Tables.** Det hele inklusive p-værdien beregnes dog i udskrifen på lommeregnerens skærm.

T - baseret test med σ ukendt

På tilsvarende måde kan man nu opstille et test for to mindre stikprøver, hvor standardafvigelsen i totalpopulationen σ ikke er kendt og derfor erstattet med den tilsvarende standardafvigelse i stikprøverne benævnt henholdsvis s_1 og s_2. Testeren følger nu en *t-fordeling* ganske som i det forrige sæt noter.

I forbindelse med formuleringen af testeren opstår der det problem, at testet kan udføres under to forskellige antagelser med hensyn til varianserne i de to stikprøver. De kan enten antages at være identiske eller forskellige. Praksis er udfaldet af de to varianter som regel identiske, men for en god ordens skyld gennemgås i det følgende begge muligheder. Senere i noterne ses der på, hvordan et test for identiske varianser kan udføres.

I begge tilfælde er hypoteserne som gennemgået under Z-testet.

Situation 1: T - baseret test med identiske varianser

Dette test er det hyppigst anvendte. Testeren bliver nu:

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - D_0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Testeren er *t-fordelt* med frihedsgrader lig med $fg = (n_1 + n_2 - 2)$. Der fratrækkes 2 frihedsgrader, da der anvendes information til at beregne middelværdierne i hvert datasæt.

s_p^2 er den "pulje variansen" (på engelsk "pooled variance"). Denne er et vægtet gennemsnit af de to stikprøvevarianser fra hvert datasæt. Den beregnes som:

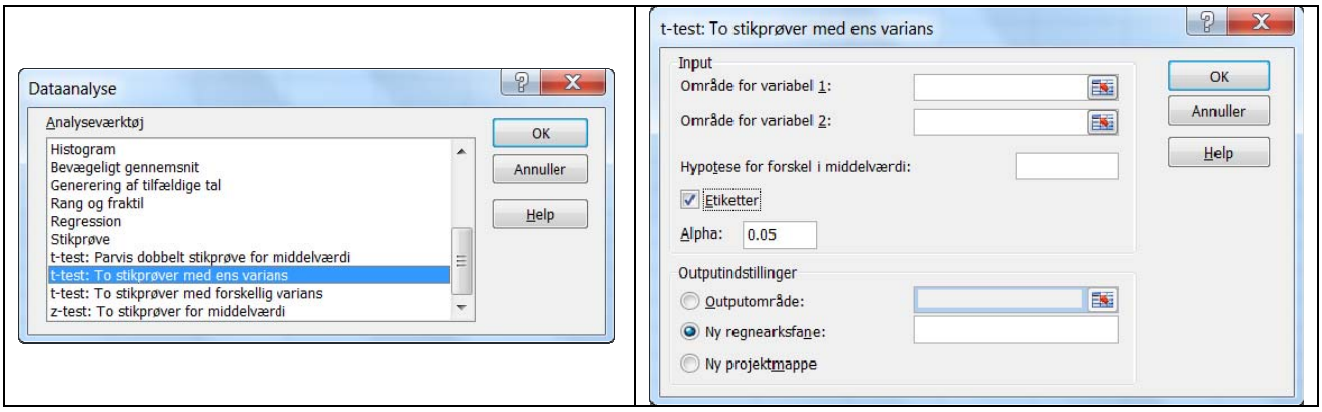
$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Et konfidensinterval for forskellen mellem de to middelværdier kan beregnes som:

$$\left[(\overline{X}_1 - \overline{X}_2) \pm t_{\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right]$$

Antallet af frihedsgrader er lig $fg = (n_1 + n_2 - 2)$. For et 95 procent konfidensinterval haves at $\alpha = 0.05$ det vil side $\alpha/2 = 0.025$.

I Excel anvendes In Excel anvendes ***Data/Data analyse/t-test: To Stikprøver med ens varians*** som vist nedenfor til venstre. Marker "OK" og så vises billedet til højre. Her markeres de indtastede dataserier.



På **TI-84 lommeregneren** tastes "stat" → "Tests" → "4: 2-SampTtest" → Stats → (nu indsættes værdierne for middelværdierne, standardafvigelserne og antallet af observationer i datasættene).

Alternativt til "stats" kan man anvende "data", og anvende de data, som man eksempelvis har indtastet i registrene *L1* og *L2*.

Man kan vælge mellem etsidet og tosidet test. Dernæst vælges om man skal anvende puljevariansen "pooled" (dette anbefales), så der svares "yes". Endelig vælges "calculate" → Nu fremkommer den beregnede t-værdi. Denne sammenlignes nu med den kritiske t-værdi fra normalfordelingen i **Statistics Tables.**

Situation 2: T - baseret test med forskellige varianser
Nu erstattes puljevariansen af stikprøvevarianserne. Testeren bliver nu:

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

Antallet af frihedsgrader fremkommer ved anvende af følgende udtryk (nok den mest bøvlede formel at arbejde med i dette kursus):

$$fg = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}$$

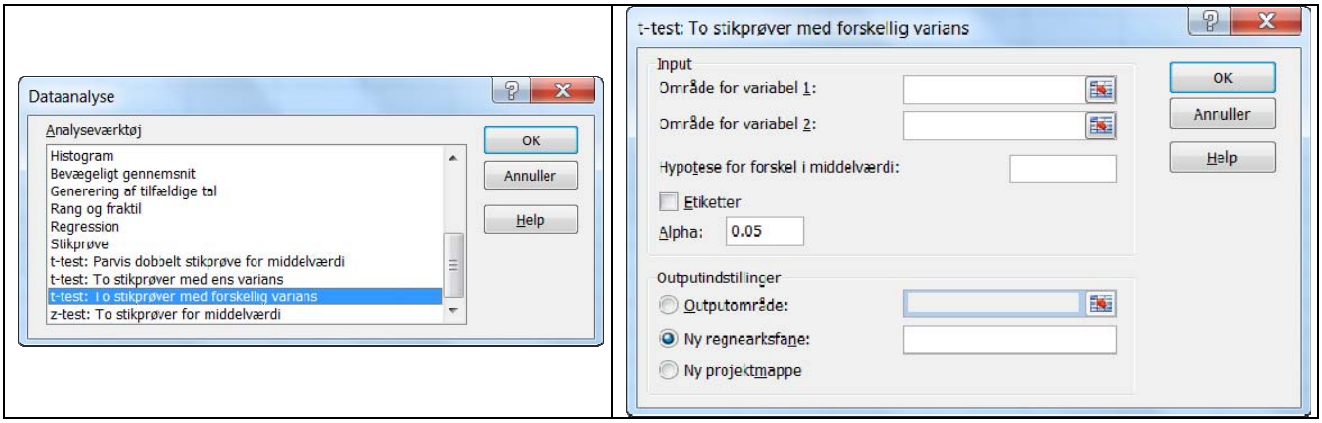
Her afrundes *fg* til det nærmeste heltal. Et konfidensinterval for forskellen mellem de to middelværdier kan beregnes som:

$$\left[(\overline{X}_1 - \overline{X}_2) \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right]$$

Antallet af frihedsgrader er givet ved udtrykket ovenfor. For et 95 procent konfidensinterval haves at $\alpha = 0.05$. Det vil sige $\alpha/2 = 0.025$.

På **lommeregneren** udføres dette test ganske som før. Det vil sige ”stat” → ”Tests” → ”4: 2-SampTtest” og så videre. Eneste forskel er, at under ”pooled” nu svares ”no”.

I Excel vælges ***Data/Data analyse/t-test: To Stikprøver med forskellig varians.*** Nu fremkommer nedenstående skærmbilleder. Arbejdsgangen er som tidligere.



2. Sammenligning af to populationsandele

Analysen kan udvides til at omfatte en sammenligning af to populationsandele. Dette er et test for totalpopulationer og datasættene antages derfor at være normalfordelte. Testet er en udvidelse af det test, der omtales i det forrige sæt noter afsnit 3, til at omfatte to populationsandele, der benævnes $\hat{p}_1$ og $\hat{p}_2$.

Opstil som et eksempel følgende tosidede hypoteser:

H_0: $\hat{p}_1 - \hat{p}_2 = 0$ (de to populationsandele er identiske)
H_1: $\hat{p}_1 - \hat{p}_2 \neq 0$ (de to populationsandele er forskellige)

Som tidligere kan der indsættes en hypotetisk forskel mellem populationsandel D_0 for en given forskel mellem $\hat{p}_1$ og $\hat{p}_2$.

Datasættene kan defineres som følger: Normalt er størrelsen af datasættene forskellig, så der vil som regel gælde at $n_1 \neq n_2$. Populationsandelene $\hat{p}$ er defineret som $\hat{p}_1 = x_1 / n_1$ og $\hat{p}_2 = x_2 / n_2$ hvor x_1 eksempelvis er antallet som har sagt "ja" i datasæt 1, mens x_2 er asntallet, som har sagt "ja" i datasæt 2.

Der kan, som i tilfældet ovenfor, beregnes en puljeandel som følger:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Testeren kan opstilles som:

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - D_0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

For at testet skal være gældende forudsættes at n_1p_1, $n_1(1-p_1)$, n_2p_2, og $n_2(1-p_2)$ alle er mindst lig med 5.

Som ovenfor kan der opstilles et konfidensinterval for forskellen mellem populationsandelene:

$$\left[(\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-p_2)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right]$$

For et 95 procent konfidensinterval haves at $\alpha = 0.05$. Det vil sige $\alpha/2 = 0.025$.

Eksempel: Bageren på Langeland – igen!
Vi har tidligere beskæftiget os med bageren på Langeland, og hans kager med ”ja” og ”nej”!

I denne afsluttende problemstilling ser vi på validiteten af den stikprøve, som bagereens antal af solgte kager kan antages at udgøre.

Ved afstemningen den 3. februar 2003 om sammenlægningen af de 3 kommuner på Langeland, var det muligt via tekst-tv på DR, at finde informationer om valgets aktuelle udfald. Ved anvendelse af denne kilde fandtes det, at der var 3,400 stemmeberettigede, hvoraf 2.695 afgav stemme. I alle 3 kommuner var der kvalificeret flertal af ”ja” til sammenlægningen, så den blev gennemført. Af de 2,695 afgivne stemme var 1,665 stemmer ”ja”.

Undersøg nu, om den stikprøve, som bagerens salg af kager udgør, var repræsentativ for valgets udfald ved et 95 procents signifikansniveau.

Fra det forrige sæt af noter erindres det (forhåbentlig), at bageren solgte 656 kager, hvoraf 385 havde påskriften ”ja”. Dette svarede til 58.7 procent. Dette salg kan betragtes som en prognose på valgets udfald.

Med disse størrelser af de 2 datasæt, kan der ikke være problemer forbundet med at anvende normalfordelingen.

Lad nu bagerens prognose være datasæt 1, og valgets aktuelle udfald være datasæt 2. Andelene af ”ja” stemmer i de to datasæt er henholdsvis:

$$\hat{p}_1 = \frac{x_1}{n_1} = \frac{385}{656} = 0.587 \qquad \text{og} \qquad \hat{p}_2 = \frac{x_2}{n_2} = \frac{1665}{2695} = 0.618$$

Så ved valget blev der afgivet 61.8 procent ”ja” stemmer. Problemstillingen er således, om de 61.8 procent adskiller sig signifikant fra de 58.7 procent, som blev resultatet af bagerens salg af kager. De analyserede hypoteser kan på denne baggrund opstilles som følger:

H_0: $\hat{p}_1 - \hat{p}_2 = 0$ (bagerens prognose er korrekt)
H_1: $\hat{p}_1 - \hat{p}_2 \neq 0$ (bagerens prognose er *ikke* korrekt)

Da bagerens prognose både kan over- og undervurdere andelen af ”ja”-stemmer, så vil et tosidet test være det mest hensigtsmæssige. Da der testes på et 95 procents niveau, så må der gælde at $\alpha=0.05$ og $\alpha/2 = 0.025$.

Indledningsvis beregnes puljeandelen som:

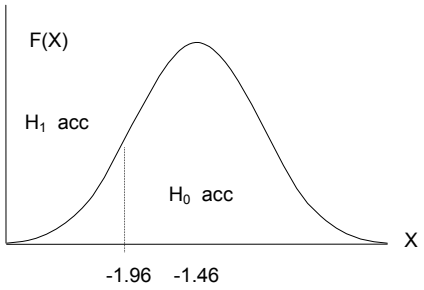
$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{385 + 1665}{656 + 2695} = \frac{2050}{3351} = 0.612$$

Testeren kan beregnes som:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.587 - 0.618}{\sqrt{0.612(1-0.612)\left(\frac{1}{656} + \frac{1}{2695}\right)}} = \frac{-0.031}{0.0212} = -1.46$$

Den kritiske værdi er lig med $Z_{\alpha/2} = Z_{0.025} = \pm 1.96$, som findes i **Statistics Tables.** Da $-1.96 < -1.46$ accepteres H_0. Det vil sige, at de 2 datasæt's udfald ikke kan siges at være signifikant forskellige. Det betyder også, at bagerens salg af kager med "ja" var en korrekt prognose for udfaldet af valget om sammenlægninger af kommuner på Langeland.

Illustration



Vi kan beregne *p-værdien* til

$$P(Z < -1.46) + P(Z > 1.46) = 2P(Z < -1.46) = 2(0.0721) = 0.1442$$

Da denne værdi er større end 0.1 er der ikke engang tale om svag signifikans.

Endelig er der forudsætningerne for testet. Her er (afrundet):

$n_1p_1 = 656(0.587) = 385$ $\quad$ $n_1(1-p_1) = 656(1-0.587) = 271$
$n_2p_2 = 2695(0.618) = 1665$ $\quad$ $n_2(1-p_2) = 2695(1-0.618) = 1029$

Da alle er over 5 er forudsætningerne opfyldt.

På **lommeregneren** løses problemstillingen ved at taste "stat" → "Tests" → "6: 2-PropZtest" → (nu indsættes værdierne for x_1, n_1 og x_2, n_2). Ved at tryukke på "calculate fremkommer det netop fundne resultater inklusive p-værdien (dog her findes til 0.1450 sandsynligvis som en følge af, at det ikke er muligt at lave et opslag på normalfordelingen, som er tilstrækkeligt præcist).

3. Sammenligning af to varianser og F-fordelingen

I det første af de to t-test, som blev gennemgået i afsnit 2, blev det antaget, at varianserne var identiske. Dette bør man undersøge validiteten af ved et test. Benævnes varianserne i totalpopulationerne 1 og 2 med henholdsvis σ_1^2 og σ_2^2, så svarer dette til at vi undersøger hypoteserne:

H_0: $\sigma_1^2 = \sigma_2^2$ (de 2 datasæt har identiske varianser)

mod

H_1: $\sigma_1^2 > \sigma_2^2$ (variansen datasæt 1 er større end i datasæt 2)

Dette benævnes også, at der under nulhypotesen gælder, at der er *homogenitet* mellem varianserne.

Et test på varianserne er fundamentalt forskelligt fra et test på middelværdien. Dette skyldes, at variansen er en kvadreret størrelse. Variansen er derfor *altid* positiv. Derfor skal den fordeling, som beskriver variansen, også være positiv. Man der derfor nødt til at introducere en kvadreret fordeling for at kunne håndtere problemstillingen

Et konfidensinterval for variansen og standardafvigelsen

Betragt først en mere forsimplet problemstilling, hvor man alene prøver at opstille et konfidensinterval for variansen i et enkelt datasæt. Betragt eksempelvis den tidligere anvendte illustration med prisen på benzin. Over 25 besøg på tankstationen fandtes, at middelværdien var 9.95 DKK med en standardafvigelse på 0.30 DKK svarende til en varians på $(0.30)^2 = 0.09 = s^2$.

Et konfidensinterval for denne varians benævnt *s*, da vi er i en stikprøve, kan skrives som:

$$\left[\frac{(n-1)s^2}{\chi^2_{\alpha/2}}; \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right]$$

Ofte ønsker man at opstille et konfidensinterval for *standardafvigelsen*. Dette gøres ved at tage kvadratroden på konfidensintervallet for variansen. Man får da:

$$\left[\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}}; \sqrt{\frac{(n-1)s^2}{\chi^2_{1-(\alpha/2)}}}\right]$$

I begge formler er χ^2 lig med **chi-i-anden fordelingen**. Det er basalt en kvadreret og transformeret normalfordeling. Denne fordeling er som normalfordelingen meget stabil, og har mange anvendelser, som det vil fremgå af de følgende sæt at noter. **Chi-i-anden** fordelingen er tabuleret i **Statistics Tables** på side 11. Her fremgår det, at chi-i-anden

fordelingen ganske som t-fordelingen afhænger af antallet af frihedsgrader. Da man allerede har beregnet middelværdien for datasættet, så er antallet af frihedsgrader lig med *fg = (n−1).*

Antages det, at man ønsker at finde et 95 procents interval for variansen eller standardafvigelsen, så er er α = 0.05 og α/2 lig med 0.025. Af formlerne fremgår, at man også skal bruge værdien for (1−α/2) = 0.975. Dette tal skal det også være muligt at finde i tabellen.

Man kan nu opstille et 95 procents konfidensinterval for standardafvigelsen på prisen på benzin. Ved indsættelse af værdierne i formlen findes:

$$\left[\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}};\sqrt{\frac{(n-1)s^2}{\chi^2_{1-(\alpha/2)}}}\right]=\left[\sqrt{\frac{(25-1)*(0.30)^2}{39.3641}};\sqrt{\frac{(25-1)*(0.30)^2}{12.4011}}\right]=[0.2342\,;0.4173]$$

Hvordan fremkommer tallene for χ^2-værdierne? Nedenfor er angivet, hvordan man finder disse tal på side 11 i **Statistics Tables.** Indledningsvis skal man finde antallet af frihedsgrader. Dette findes til *fg = n−1 = 25−1 = 24.*

Dernæst går man ind i tabellen under *24* i forspalten og finder værdierne for 0.025 og 0.975 i hovedspalten. Derved fremkommer to værdier for chi-i-anden, som er vist med signatur i tabellen.

	χ values							
df	**0.99**	**0.975**	**0.95**	**0.90**	**0.10**	**0.05**	**0.025**	**0.01**
1	0.0002	0.0010	0.0039	0.0158	2.7055	3.8415	5.0239	6.6349
…	…	…	…	…	…	…	…	…
24 →	10.8564	12.4012	13.8484	15.6587	33.1962	36.4150	39.3641	42.9798
…	…	…	…	…	…	…	…	…
26	12.1981	13.8439	15.3792	17.2919	35.5632	38.8851	41.9232	45.6417

Det ses således, at

$$\chi^2_{0.025;24}=39.3641\,og\;\chi^2_{0.975;24}=12.4011$$

Værdierne indsættes i formlen ovenfor. Bemærk at det største tal skal stå til *venstre.* Det bliver da den største divisor, og man får da her det mindste tal, som er den nedre grænse for konfidensintervallet.

Test for identiske varianser
Med denne viden om konfidensintervallet for en standardafvigelse eller varians kan man gå tilbage til den oprindelige problemstilling og udvikle et test for identiske varianser.

Husk fra indledningen på side 10, at hvis varianserne var identiske, så havde man *homogenitet.* Denne viden kan man benytte! Hvis varianserne er identiske, så er de lige store. Det betyder, at ved division af den ene med den anden varians, så fås 1.

Divideres hypoteserne på side 10 øverst således σ_2^2 fås følgende hypoteser:

$$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1 \quad \text{og alternativet} \qquad H_1: \frac{\sigma_1^2}{\sigma_2^2} > 1$$

Så hvis forholdet mellem varianserne er tilstrækkeligt større end ét, så vil man afvise H_0 og acceptere H_1.

Da vi opstillede konfidensintervallet for variansen, anvendtes chi-i-anden fordelingen. Nu ser man på forholdet mellem 2 varianser, der repræsenterer 2 datasæt, som hver er chi-i-anden-fordelt. Dette forhold beskrives i **F-fordelingen**, som findes i **Statistics Tabeles** på siderne 12 til 14. Fordelingen er opkaldt efter den engelske forsker Sir Ronald A. Fisher, som udviklede denne fordeling i 1920'erne (til plage for enhver studerende, der har statistik)!

Da chi-i-anden-fordelingen har frihedsgrader, gælder dette også for F-fordelingen. Kald frihedsgraderne for henholdsvis fg_1 og fg_2, da vi fordelingen for eksempelvis $\alpha = 0.025$ se ud som på side 13 i **Statistics Tables.**

Table of F-Values 0,025

This Table was generated by use of the Excel function FINV

fg 2 ↓ \ fg 1 →	1	…	9	10	15	…	∞
1	648	…	963	969	985	…	1018
2	38.51	…	39.39	39.40	39.43	…	39.50
3	…	..	…	…	…	…	…
9	7.21	5.71	4.03	3.96	***3.77***	…	3.33
10	6.94	5.46	3.78	3.72	3.52	…	3.08
15	6.20	4.77	3.12	3.06	2.86	…	2.40
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
∞	5.03	3.69	2.12	2.05	1.84	1.21	1.00

Bemærk nederst i højre hjørne. Går antallet af frihedsgrader mod uendelig for begge datasæt fås at den kritiske F-værdi er lig med 1. Dette er præcis, som forventet, når man ser på nulhypotesen, som formuleret øverst på den foregående side. Her er det netop forholdet mellem varianserne, der er i fokus. F-værdierne angiver således hvor stor en afvigelse, der er tilladt mellem varianserne ved givne størrelser af data.

Betragt det fremhævede eksempel i tabellen. Vi ser på to datasæt med henholdsvis 11 observationer i datasæt 1 og 16 observationer i datasæt 2. Da er frihedsgraderne lig med $fg_1 = (n_1-1) = 11-1 = 10$ og $fg_2 = (n_2-1) = 16-1 = 15$.

Man kan nu finde F-værdien ved at aflæse for fg_1 (læses i hovedspalten eller vandret) til at være lig med 10 og fg_2 (læses i forspalten eller lodret) til at være lig med 15. Dette skrives som $F_{0.025}(10,15) = 3.06$.

Hvad betyder dette? Med den givne usikkerhed tillades variansen i datasæt 1 at være godt 3 gange så stor som i datasæt 2 og stadig acceptere H_0. Dette lyder af meget, men man skal huske på, at datasættene er meget små. Bemærk, som nævnt ovenfor, at denne tolerance falder, når datasættene bliver større.

Ved opstilling af F-testet for identiske varianser er der to forhold, som er vigtigere. For det første, placerer man altid, den største varians i tælleren. Derved antager testeren altid en værdi, der er større end ét. For det andet udføres som en ensidet test ved en α-værdi på 0.05 og som et tosidet test ved en α-værdi på 0.025.

Eksempel
Betragt et lille eksempel. Vi har to datasæt med henholdsvis $n_1=16$ og $n_2=10$ observationer. I det første datasæt er variansen lig med 3.8, mens den i det andet datasæt er lig med 1.6. Antag eksempelvis et tosidet test. Er varianserne identiske ved et signifikansniveau på $\alpha = 0.025$?

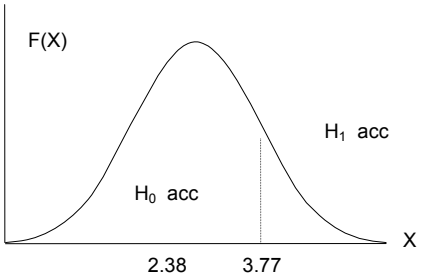
Her findes først forholdet mellem varianserne under H_0. Det er lig med:

$$H_0: \frac{s_1^2}{s_2^2} = \frac{3.8}{1.6} = 2.38$$

Den kritiske værdi findes nu i fordelingen, idet vi antager at $\alpha = 0.025$. Frihedsgraderne er lig med $fg_1 = (n_1-1)=(16-1) = 15$ og $fg_2 = (n_2-1) = (10-1) = 9$. Dette skrives som $F_{0.025}(15,9)$. I tabellen ovenfor findes den kritiske værdi til at være lig med **3.77**.

Da $2.38 < 3.77$ accepteres H_0, og de to varianser er identiske givet de meget små datasæt.

Illustration



Som det ses, er F-testet oftest et ensidet test jævnfør også formuleringen af hypoteserne på side 12.

Hvad nu hvis variansen i datasæt 1 er *mindre* end i datasæt 2? Så vender man, som nævnt på forrige side, testet om, således at forholdet mellem varianserne *altid* bliver større end ét.

Løsning med en F-værdi, der er mindre end ét, er muligt, men så skal man vende fordelingen om. Dette falder udenfor rammerne af dette kursus!

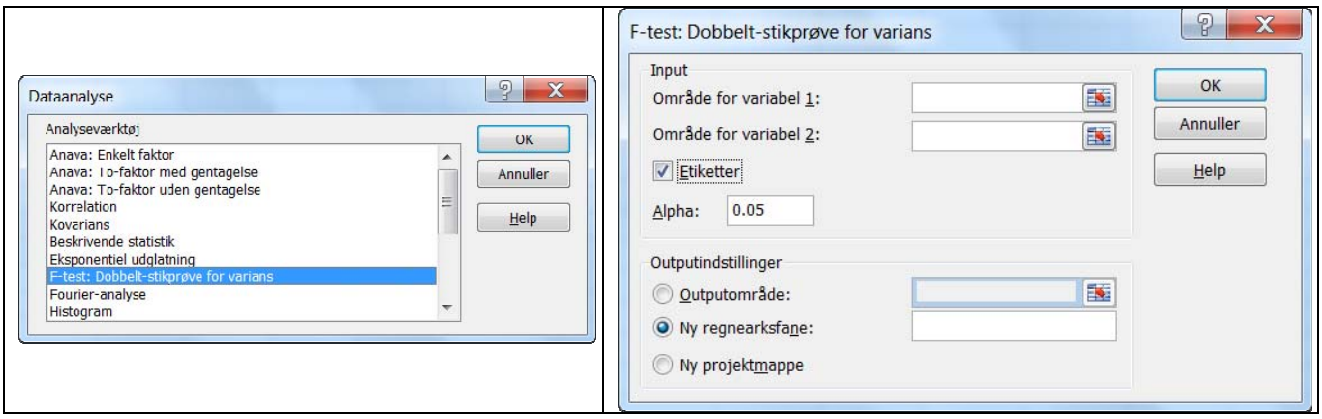
Hvordan går man nu på **lommeregneren?** tastes "stat" → "Tests" → "E: 2-SampFtest" → Stats → (*standardafvigelserne* (i dette tilfælde er $\sqrt{3.8}=1.95$ og $\sqrt{1.6}=1.26$). Endvidere skal antallet af observationer i datasættene angives).

Med hensyn til hypoteser, så markeres $<\sigma_2$ (testet udføres som ensidet test)

I udskriften findes (som forventet), at F = 2.395 (bedre afrunding), mens der findes en *p-værdi* på 0.094.

Da p-værdien er større end 0.025, accepteres H_0 ganske som ovenfor.

I Excel vælges ***Data/Data analyse/F-test: Dobbelt stikprøve for varians.*** Nu fremkommer nedenstående skærmbilleder. Arbejdsgangen er som tidligere.



5. Gennemarbejdede eksempel

Effekten af brug af en ny type motorolie[4]

Et kendt bilmærke har solgt mange biler med en meget populær dieselmotor. Denne motor har serviceinterval for hver 30.000 kilometer.

Et serviceværksted foreslår en kunde, at man kan skifte til en mere fin motorolie. Denne olie vil holde kortere tid, men dette er ikke af betydning, hvis serviceintervallet er kortere. Den finere olie vil imidlertid resultere i reduceret friktion i motoren, og dermed en forbedret brændstoføkonomi. På denne måde vil der kunne opnås en besparelse på brændstof.

Kunden, som fører kørselsregnskab, har undersøgt dette forhold i to perioder med identiske klimaforhold. Nedenstående tabeller viser det gennemsnitlige antal kørte kilometer per liter henholdsvis før (benævnt F) og efter (benævnt E) skift til den finere motorolie. I stikprøven "før" er der 12 observationer, mens der i stikprøven "efter" er 14 observationer.

Før	16.6	17.5	16.8	17.2	15.1	16.1	15.8	16.3	16.1	15.8	16.3	17.2

Efter	17.8	17.5	16.9	18.1	17.8	16.6	17.2	16.9	17.5	19.2	17.2	18.1	18.8	16.3

Spørgsmål:

A. Beregn middelværdi, varians og standardafvigelse for de to datasæt
B. Undersøg en hypotese om, at den nye olie har forbedret brændstoføkonomien
C. Undersøg antagelsen om, at en identisk varians er opfyldt mellem de to datasæt

Det første spørgsmål er repetition fra Statistik I, men medtages for en god ordens skyld.

Til løsningen kan anvendes enten Excel eller lommeregneren.

A)

Beregning af middelværdi og standardafvigelsen

Før ny olie: $\overline{X}_F = \frac{\sum_{i=1}^{n_F} x_{F.i}}{n_F} = \frac{196.8}{12} = 16.40$ km per liter benzin

$$s_F = \sqrt{\frac{\sum_{i=1}^{n_F} (x_{F.i} - 16.40)^2}{n_F - 1}} = 0.6941$$

[4] Opgaven er taget fra skriftlig eksamen på HA-studiet januar 2009. Data hidrører fra det virkelige liv!

Resultatet for standardafvigelsen kunne også være opnået ved anvendelse af følgende formel fra Statistik I kurset:

$$s_F = \sqrt{\frac{1}{n_F - 1}\left[\sum_{i=1}^{n_{Før}} x_{F.i}^2 - \frac{(\sum_{i=1}^{n_{Før}} x_{F.i})^2}{n_F}\right]} = \sqrt{\frac{1}{12-1}\left[3{,}232.82 - \frac{(196.8)^2}{12}\right]} = 0.6941$$

Nu beregnes for det andet datasæt:

Efter ny olie: $\overline{X}_E = \frac{\sum_{i=1}^{n_E} x_{E.i}}{n_E} = \frac{245.9}{14} = 17.56$ km per liter benzin

$$s_E = \sqrt{\frac{\sum_{i=1}^{n_E}(x_{E.i} - 17.56)^2}{n_E - 1}} = 0.8120$$

Eller alternativt:

$$s_E = \sqrt{\frac{1}{n_E - 1}\left[\sum_{i=1}^{n_A} x_{E.i}^2 - \frac{(\sum_{i=1}^{n_E} x_{E.i})^2}{n_E}\right]} = \sqrt{\frac{1}{14-1}\left[4{,}327.63 - \frac{(245.9)^2}{14}\right]} = 0.8120$$

B)
Undersøg en hypotese om, at den nye olie har forbedret brændstoføkonomien

Der anvendes et t-test til at undersøge problemstillingen. For lethedens skyld, antages der identiske varianser. Der undersøges for om varianserne er identiske i det afsluttende spørgsmål i opgaven.

De det er interessant om motorolien får bilen til at køre længere på en liter benzin opstilles der en ensidet test til belysning af problemstillingen.

Følgende hypoteser formuleres:

H_0: $\mu_E \leq \mu_F$	Den nye olie har igen signifikant effekt
H_1: $\mu_E > \mu_F$	Den nye olie får bilen til at køre længere på literen

Nu beregnes puljevariansen som:

$$S_p^2 = \frac{(n_E - 1)s_E^2 + (n_F - 1)s_F^2}{n_E + n_F - 2} = \frac{(14-1)(0.8120)^2 + (12-1)(0.6941)^2}{14 + 12 - 2} = \frac{8.5715 + 5.2995}{24} = 0.5779$$

Testeren kan beregnes til:

$$t = \frac{\overline{X}_E - \overline{X}_F}{\sqrt{s_p^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} = \frac{17.56 - 16.40}{\sqrt{0.5779\left(\frac{1}{14} + \frac{1}{12}\right)}} = \frac{1.16}{0.2990} = 3.89$$

Antallet af frihedsgrader er lig med $fg = n_E + n_F - 2 = 14 + 12 - 2 = 24$

Den kritiske værdi for testet findes ved anvendelse af side 10 **Statistics Tables.** Vi har antaget, at testet er ensidet, så $\alpha = 0.05$. Værdien er da lig med *1.711*.

Da $1.711 < 3.89$ forkastes H_0. Det vil sige, at den nye motorolie får bilen til at køre længere på literen.

Ved anvendelse af Excel kan man udføre beregningen på tilsvarende vis. Anvend ***Data/Data analyse/t-test: To Stikprøver med ens varians*** som beskrevet ovenfor.

t-test: To stikprøver med ens varians

	Efter	*Før*
Middelværdi	17.56	16.40
Varians	0.66	0.48
Observationer	14	12
Puljevarians	0.58	
Hypotese for forskel i middelværdi	0	
fg	24	
t-stat	3.89	
P(T<=t) en-halet	0.00	
t-kritisk en-halet	1.71	
P(T<=t) to-halet	0.00	
t-kritisk to-halet	2.06	

Udskriften både resultatet af det ensidede og det tosidede test. I det tosidede tilfælde findes t-værdien til 2.06. Her er $\alpha = 0.025$. I begge tilfældene er p-værdierne langt under 0.05.

Ved anvendelse af **lommeregneren** skal man huske, at det er standardafvigelserne, som skal angives. Endvidere skal man i udskriften på lommeregneren være opmærksom på, at det

ikke er puljevariansen, men puljestandardafvigelsen der fremkommer. Denne er lig med $\sqrt{s_p^2} = \sqrt{0.5779} = 0.7602$

C)
Det skal nu undersøges, om de to varianser er identiske.

Der er to datasæt med henholdsvis $n_F = 12$ og $n_E = 14$.

Indledningsvis beregnes de to varianser ved at kvadrere standardafvigelserne fra spørgsmål A:

$$s_F = 0.6941 \Rightarrow s_F^2 = 0.4818$$
$$s_E = 0.8120 \Rightarrow s_E^2 = 0.6593$$

Da variansen for *E* er den største bruges denne værdi i tælleren i det følgende test.

Der opstilles følgende hypoteser:

H_0: $\sigma_E^2 = \sigma_F^2$ (varianserne for E og F er identiske)

mod

H_1: $\sigma_E^2 > \sigma_F^2$ (variansen E er større end F)

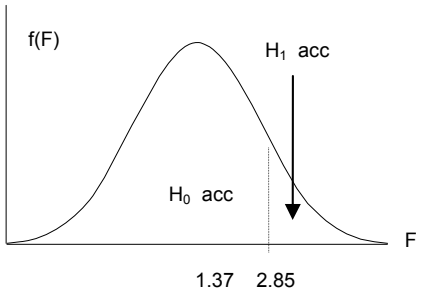
Tester:

$$F_{\alpha(n_E-1)(n_F-1)} = \frac{s_E^2}{s_F^2} = \frac{(0.8120)^2}{(0.6941)^2} = \frac{0.6593}{0.4818} = 1.37$$

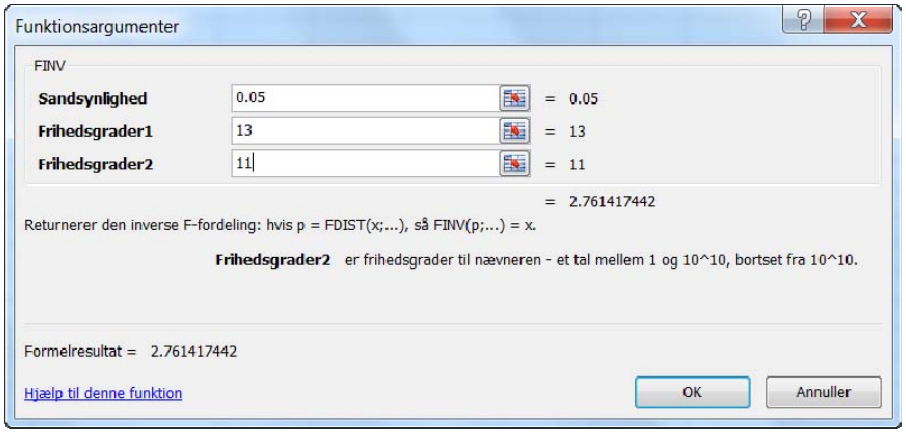
For nærværende er der udført et ensidet test med α = 0.05. Ønskes testet tosidet vælges α = 0.025. I **Statistics Tables** kan man finde den kritiske værdi som $F_{0.05(13,11)} = 2.85$. (Værdien her er gengivet for $F_{0.05(15,10)}$ som en følge af manglende frihedsgrader i tabellen).

Da 1.37 < 2.85 accepteres H_0. Det vil sige, at varianserne i de to stikprøver er identiske. Forudsætningen er således opfyldt for at kunne anvende det test, der blev brugt i spørgsmål B. for det i spørgsmål B.

Illustration:



Den præcise værdi kan findes enten i Excel med ***Formler/indsæt funktion/Statistik/FINV*** som vist i skærmbilledet nedenfor. Her findes den præcise værdi til 2.76



Testet kan også udføres i Excel. Anvend ***Data/Data analyse/F-test: Dobbelt stikprøve for varians*** som beskrevet ovenfor.

Følgende udskrift fremkommer:

F-test: Dobbelt stikprøve for varians

	Efter	*Før*
Middelværdi	17.56	16.40
Varians	0.66	0.48
Observationer	14	12
fg	13	11
F	1.37	
P(F<=f) en-halet	0.30	
F-kritisk en-halet	2.76	

Som forventet!

På **lommeregneren** skal man huske, at der bedes om standardafvigelsen.

Nils Karl Sørensen ©
Statistik II, BA int, Flensborg

Sæt 3: Ensidet variansanalyse (ANOVA)

af Nils Karl Sørensen

Indhold *side*

1. Hvad er variansanalyse?

I de to foregående sæt af noter har vi først beskæftiget os med at undersøge validiteten af en given påstand i forhold til et givet datasæt. Dernæst blev problemstillingen udvidet til at omfatte en sammenligning af datasæt, hvor vi dels så på middelværdien eller populationsandelen dels på varianserne eller standardafvigelserne.

I dette sæt af noter generaliseres problemstillingen til at omfatte en sammenligning af p datasæt for identisk middelværdi. Ordet "variansanalyse" er således lidt misvisende i forhold til den analyserede problemstilling. Med hensyn til den anvendte metode er ordret dog ganske passende. Rent teknisk kan man nemlig ikke håndtere p grupper. Derfor foretages en dekomponering af datasættene efter to kriterier. Disse kriterier sammenholdes dernæst ved at sammenligne varianserne for hvert kriterium. Deraf navnet *variansanalyse*. På engelsk ANOVA (ANalysis Of VAriance).

Når man taler om *ensidet variansanalyse*, så refereres der en til analyse af en enkelt variabel, hvor man undersøger p datasæt for identiske middelværdier. Det vil sige, at der er én faktor eller variabel. Det kan eksempelvis være, at man undersøger prisen for en række produkter i en række forskellige supermarkeder i Flensborg, Sønderborg og Aabenraa.

Udvider man problemstillingen til at sige, at der i hver by er tre forskellige supermarkeder, som udbyder produkterne, så har man det der kaldes *tosidet variansanalyse.* Supermarkederne kan eksempelvis være Aldi, Lidl, Netto og Coop/Rewe. *Tosidet*

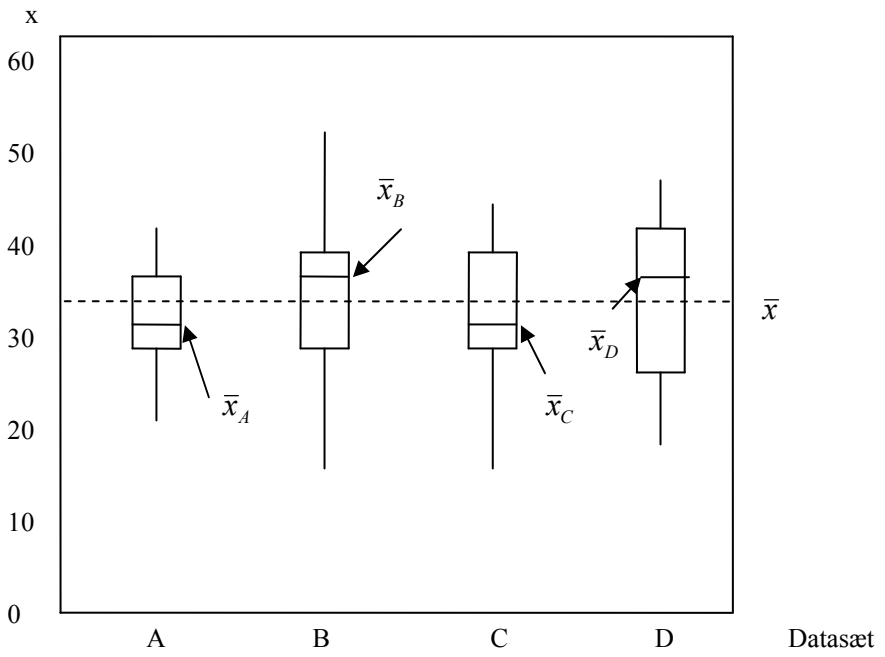
variansanalyse er ikke en del af dette kursus, men danner udgangspunkt for *faktoranalyse,* der ofte finder anvendelse indenfor analyse af spørgeskemaer i forbindelse med marketing.

Eksempler på anvendelse af ensidet variansanalyse kan være:

- Sammenligning af p salgsdistrikter for at finde, hvor det gennemsnitlige salg eller omsætningen er størst
- Sammenligning af salget på måneder i p år
- Undersøgelse af salgseffekten af annoncekampagner i p områder
- Undersøgelse af kvartalsvise sæsonmønstre (som det vises i afsnit 5 i disse noter)

Og meget mere!

Betragt, som en illustration, diagrammet nedenfor. Her antages 4 populationer eller grupper kaldet A, B, C og D; det vil sige $p = 4$. Det kan eksempelvis være en virksomheds omsætning på 4 forskellige eksportmarkeder. Grupperne behøver ikke at have identiske størrelser, så antallet af observationer kan variere fra gruppe til gruppe. For hver gruppe er der i diagrammet vist et boks-diagram. Den vandrette streg i boks-diagrammet angiver middelværdien i den givne gruppe.



Som det ses, er der et mønster: Middelværdierne i grupperne B og D er større end i grupperne A og C. Problemstillingen er nu, om denne forskel er signifikant? Nok er middelværdien for B og D større, men som det ses, så er spredningen af observationerne også stor. Har det en betydning? For at lokalisere det samlede materiale, kan man beregne $\bar{x}$, som er middelværdien for alle de 4 grupper under ét. Denne middelværdi kan anvendes som reference og kaldes den *samlede middelværdi*. Undersøgelsen er da et test på, om de

enkelte gruppers middelværdier er signifikant forskellige i forhold til den samlede middelværdi (på engelsk kaldet "grand mean").

Den samlede middelværdi $\bar{x}$ danner således udgangspunkt for analysen. Det særlige ved den samlede middelværdi er, at denne relaterer de fire grupper til hinanden. Man kan sige, at den samlede middelværdi er et udtryk for den horisontale integration mellem de 4 grupper. Vertikalt er der imidlertid ingen integration. Her ses der alene på spredningen indenfor grupperne.

Med udgangspunkt i den horisontale og vertikale måde at anskue illustrationen overfor på, kan man formulere to typer af variation:

- *Variation mellem grupperne:*
  Her refererer en given observation til den samlede middelværdi $\bar{x}$

- *Variation indenfor grupperne:*
  Her refererer en given observation til gruppen egen middelværdi $\bar{x}_p$

Tankegangen i ANOVA er at relatere disse to typer af variation til hinanden. Det vil også sige, at vores *p* grupper bliver dekomponeret til to typer af variation eller kriterier. I en analytisk sammenhæng er dette mest hensigtsmæssigt. Ellers skulle man sammenligne alle de *p* grupper to og to, som i det sidste sæt af noter. Det vil give en masse ekstra arbejde. Da variation både kan være positiv og negativ, må al variation kvadreres ganske som det er tilfældet, når man beregner variansen. Det er derfor, at ANOVA kaldes variansanalyse, selvom der testes for middelværdier. Der udledes kvadrater for hver type af variation, der efterfølgende testes overfor hinanden. Da der testes på to kvadrater, må testet nødvendigvis være et F-test.

Forudsætningerne for ANOVA

Med denne intuitive ballast kan man nu udlede testet. Indledningsvis opstilles antagelserne for ANOVA:

1. Der antages konstant varians (homogenitet) mellem de *p* grupper: $\sigma_1^2 = \sigma_2^2 = ... = \sigma_p^2$
2. Det antages, at alle data er normalfordelte
3. Det antages, at der er uafhængighed mellem grupperne

Den sidste antagelse betyder for eksempel, at salget i ét salgsdistrikt ikke påvirket salget i et andet salgsdistrikt.

2. Ensidet variansanalyse

For hver af de *p* grupper kan man beregne gruppens middelværdi μ_i, *i = 1,2,...,p.* Man kan nu opstille hypoteserne, som skal undersøges.

Hypoteserne i variansanalyse (ANOVA)

H_0: $\mu_1 = \mu_2 = ... = \mu_p$ (middelværdierne er identiske mellem grupperne)

H_1: Minimum én middelværdi er forskellig fra de øvrige

Hver gruppe har n_i observationer. Som tidligere nævnt kan grupperne have forskellig størrelse. Det samlede antal af observationer er lig med $n = n_1+n_2+...+n_p$.

Betragt en given observation *j* i gruppen *i,* og kald denne for x_{ij}. Denne observation kan indgå i begge de typer af variation, som blev defineret på side 3. Variationen mellem grupperne benævnes *SSTR,* mens variationen indenfor grupperne benævnes *SSE*. Endelig kan al variationen lægges sammen. Den samlede variation benævnes *SST*, og er summen af *SSTR* og *SSE*.

Variationen findes nu i forhold til den relevante middelværdi. Det er for variationen mellem grupperne det samlede gennemsnit $\bar{x}$, mens det indenfor grupperne er variationen i de enkelte grupper $\bar{x}_p$. Al variation kvadreres for at undgå, at positiv og negativ variation går ud mod hinanden.

Overvejelserne om variation kan sammenfattes som følger:

Total variation	= mellem grupper	+ indenfor grupper
Sum Square Total	= Sum Square Treatment	+ Sum Square Error

Eller:

$$SST = SSTR + SSE$$

På formel kan dette opskrives som:

$$\sum_{i=1}^{p}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})^2 = \sum_{i=1}^{p} n_i(\bar{x}_i-\bar{x})^2 + \sum_{i=1}^{p}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2$$

Bemærk at gruppernes middelværdi $\bar{x}_i$ optræder med modsat fortegn på højresiden. Det vil sige, at denne værdi går ud.

Hvad menes der med "treatment" og "error"? Det vendes der tilbage til i notesættet om regression. Groft sagt er "treatment", den variation, som ønskes, mens "error" er den

uønskede variation (indenfor grupperne), der ikke kan forklares af sammenhængen mellem grupperne.

Udførsel af test og ANOVA-tabellen

Efter at hypoteserne er opstillet, og dekomponeringen er beskrevet, kan testet opstilles. Først beregnes *SSTR* og *SSE*. Dernæst summeres disse størrelser til *SST*.

Da alle værdier er kvadrater, bliver det nogle meget store værdier. For at gøre dem mindre divideres der med antallet af frihedsgrader; ganske som det var tilfældet, da variansen blevet beregnet under *beskrivende statistik* i efterårssemestret.

Mellem grupperne *p* fragår der én frihedsgrad, der blev anvendt til at beregne den samlede middelværdi for hele datasættet. Indenfor grupperne anvendes alle *n* observationer. Her fragår de *p* middelværdier, der er beregnet for hver af grupperne.

Ved divisionen beregnes *middelkvadratsummerne*[5], MS. Der er to middelkvadratsummer, dels middelkvadratsummen mellem grupperne kaldet *MSR,* dels middelkvadratsummen indenfor grupperne kaldet *MSE*. Disse er givet som:

Mellem grupper: $MSR = \frac{SSTR}{p-1}$ (for *p* grupper minus samlet middelværdi)

Indenfor grupper: $MSE = \frac{SSE}{n-p}$ (for *n* minus *p* grupper)

Endelig kan man finde testeren ved at dividere de to middelkvadratsummer:

$F = \frac{MSR}{MSE}$ med frihedsgrader lig $fg=(p-1);(n-p)$

Alle beregningerne kan nu sammenfattes i *ANOVA-tabellen:*

Variation	*Kvadrat sum (SS)*	*Friheds-grader (fg)*	*Middelkvadratsum (MS)*	*F-værdi*
Mellem grupper	$SSTR = \sum_{i=1}^{p} n_i(\bar{x}_i - \bar{x})^2$	p − 1	MSR = SSTR/(p−1)	$F = \frac{MSR}{MSE}$
Indenfor grupper	$SSE = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$	n − p	MSE = SSE/ (n−p)	
Total	$SST = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2$	n − 1		

[5] Dette er faktisk det samme som en varians.

Som sagt ovenfor, så er den fundne "F-værdi" F-fordelt med frihedsgrader lig med (p−1);(n−p). Det totale antal observationer er lig med $n = n_1+n_2+...+n_p$.

Bemærk hvor nydeligt alle elementerne i tabellen passer sammen! Det betyder også, at hvis man mangler nogle informationer, så kan disse næsten altid beregnes ud fra den information, der er tilgængelig. Bemærk endelig, at det samlede antal frihedsgrader er lig med $n-1$, da p'erne går ud mod hinanden.

3. Supplerende analyser med konfidensintervaller

Nu er testet opstillet og udført! Er vi så glade? Måske – men ikke helt! Hvis H_0 hypotesen er forkastet, så er det fundet, at middelværdierne er forskellige. Hvis man er interesseret i at finde ud af, *hvilke(n)* af middelværdierne, som adskiller sig fra den/de øvrige, vil det være nødvendigt at foretage en supplerende analyse (på engelsk "post-hoc" analysis).

Den supplerende analyse kan udføres på flere måder, men det mest enkle er, at opstille et konfidensinterval for hver gruppe og dernæst sammenligne gruppernes konfidensintervaller. Den gruppe, hvor konfidensintervallet *ikke* overlapper med de øvrige grupper, har den middelværdi, der signifikant afviger fra de øvrige grupper.

Et konfidensinterval til at sammenligne grupperne i og h kan opstilles som:

$$\left[(\bar{x}_i - \bar{x}_h) \pm t_{\alpha/2} \sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_h}\right)} \right] \qquad \text{med frihedsgrader } fg = n-p$$

I dette tilfælde beregnes alle parvise konfidensintervaller. Mere enkelt er det, at opstille et konfidensinterval for hver af de p middelværdier, og så sammenligne:

$$\left[\bar{x}_i \pm t_{\alpha/2} \sqrt{\frac{MSE}{n_i}} \right] \qquad \text{med frihedsgrader } fg = n-p$$

4. ANOVA på lommeregneren og i Excel

Den ensidede variansanalyse er let at udføre såvel i Excel som på lommeregneren TI-84/89.

Indledningsvis tastes de p datasæt ind i **lommeregnerens** register. Haves eksempelvis 3 datasæt, så anvend *L1, L2* og *L3*. Nu vælges: STAT → TESTS → H: ANOVA(→ ENTER.

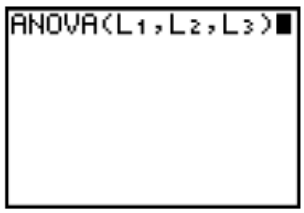
Formatet er da ANOVA(*L1,L2,L3*) → ENTER. Se følgende eksempel:

ANOVA(*liste1*,*liste2*[,...,*liste20*]**)**

I eksemplet:

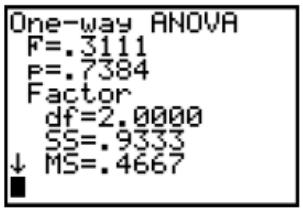
L1={7 4 6 6 5}
L2={6 5 5 8 7}
L3={4 7 6 7 6}

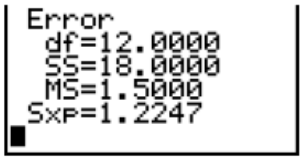
Input:



Dernæst fås følgende skærmbilleder:

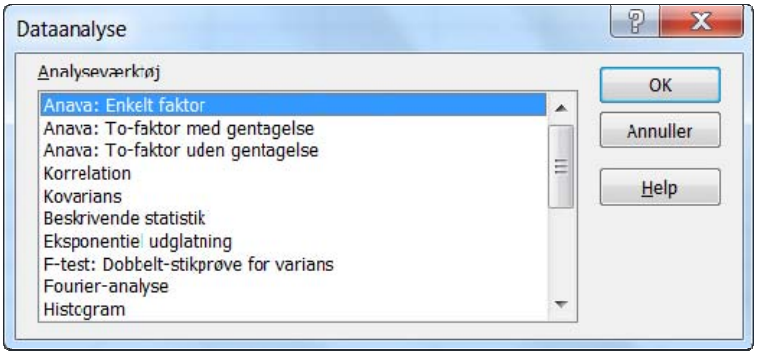
Beregnede resultater:



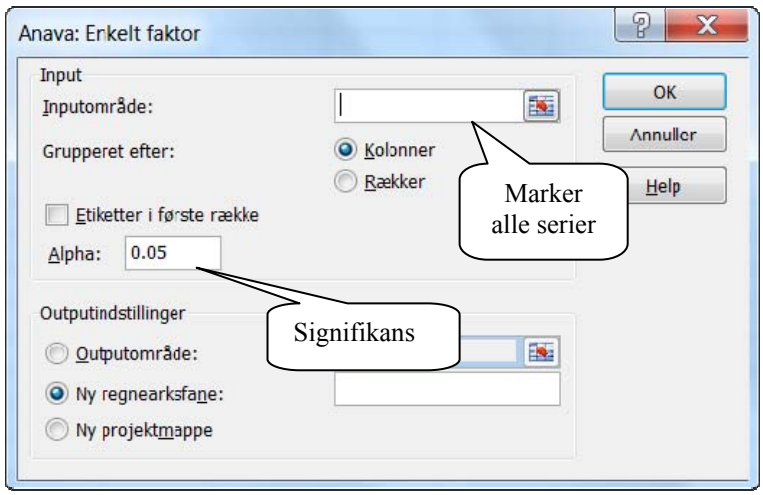


Bemærk at **SS** er summen af kvadrater og **MS** er middelkvadratet.

I **Excel** vælges ***data / data analyse / Anava: Enkelt Faktor***



Ved klik på "OK" fremkommer omstående dialogboks:



Såvel for lommeregneren som for Excel gælder, at man selv skal udføre den supplerende analyse.

5. Gennemarbejdet eksempel

Det følgende eksempel er taget fra skriftlig eksamen i statistik på HA-studiet januar 2007. Opgaven er dog tilpasset disse noter. Baggrunden for opgaven er som følger.

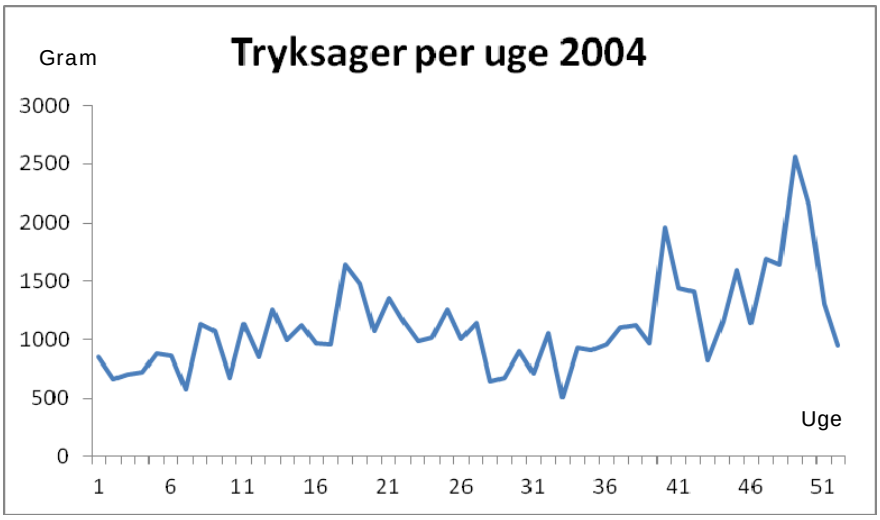
Studier af sæsonvariationen i kvartalsvise nationalregnskabsdata viser bemærkelsesværdige identiske mønstre for næsten alle OECD lande for mange variabler[6]. Generelt gælder, at der er lav økonomisk aktivitet i første og tredje kvartal, mens aktiviteten er højere i andet og fjerde kvartal. Den større aktivitet i andet kvartal skyldes, at folk her køber deres ferier, mens aktiviteten i fjerde kan henføres til julen.

I Danmark husstandsomdeles der reklamer i stigende grad hver weekend. Ofte kommer der mellem 10 og 15 reklametryksager. *Det kan meget vel tænkes at mængden af reklamer også følger det skitserede sæsonmønster.*

For at undersøge denne problemstilling indsamlede forfatteren til disse noter ugentligt reklamer for året 2004. Husholdning kan siges at være en tilfældig valgt husstand i Danmark. Året blev opdelt i 52 uger svarende til 4 kvartaler á 13 uger. For hver uge blev antallet af reklamer optalt og vejet på en digitalvægt. Det er de sidstnævnte data, der danner grundlag for denne undersøgelse. Det statistiske materiale er vedlagt som **bilag** til dette afsnit.

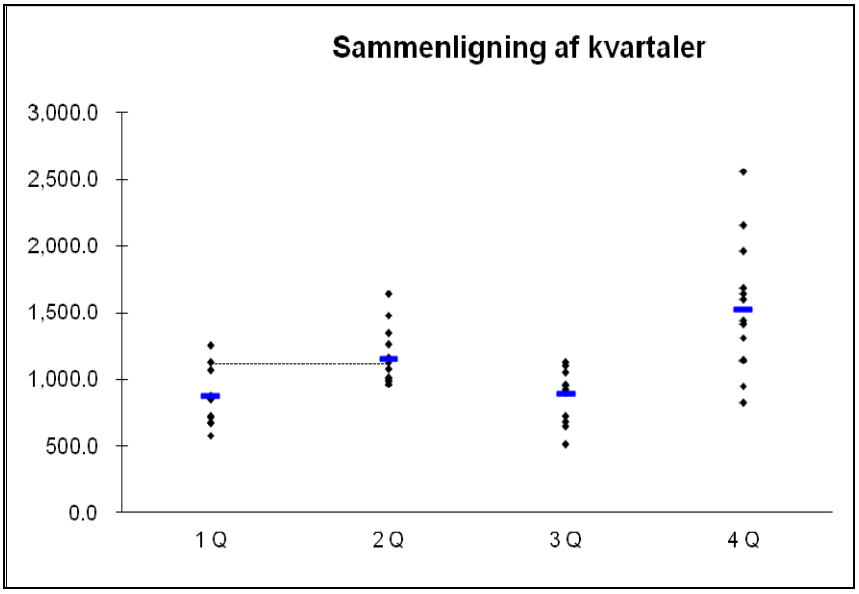
[6] Forfatteren til disse noter skrev PhD-afhandling om blandt andet dette emne. Resultaterne blev publiceret i artiklen: Svend Hylleberg, Nils Karl Sørensen and Clara Jørgensen "Seasonality in Macroeconomic Time Series". *Studies of Empirical Economics 18,* pages 321−335, 1993. Det fagligt tilfredsstillende ved dette eksempel er, at dataene på mikroniveau i denne opgave efterviser de makroøkonomiske resultater, der er præsenteret i artiklen.

For at danne sig et indtryk af problemstillingen kan man betragte nedenstående illustration, hvor der er lavet nogle grafiske illustrationer af dataene i bilaget:



Her er data vist i forhold til ugens nummer. Det fremgår, at der var mange reklamer især i ugerne omkring Påsken 2004 (ugerne 16 til 18), omkring efterårsferien (uge 40 og 41) samt i ugerne op til jul (ugerne 48 til 50).

Dette underbygger en hypotese om, at der er et sæsonmønster. Dette forhold er yderligere illustreret i nedenstående figur, der anskuer materialet lidt anderledes.



I figuren, som jeg har tegnet med et statistikprogram, der ligger udenfor rammerne af Statistik II, er data opstillet på samme måde, som i illustrationen på side 2 i nærværende sæt af noter. Det fremgår tydeligt, at middelværdien for den ugentlige mængde af reklamer er større i fjerde kvartal end for de øvrige kvartaler. Tillige ses det, at spredningen også er

større. For de øvrige kvartaler genfindes, det mønster, der er fundet på OECD-tallene. Spørgsmålet er, om variationen i middelværdierne er stor nok til at være signifikant i forhold til den samlede middelværdi, der er markeret med den vandrette linje i figuren.

Først opstilles hypoteserne. Hvis der ikke er en sæsoneffekt, vil middelværdierne være identiske, mens at mindst én (forhåbentlig fjerde kvartal) vil være forskellig fra de andre, hvis der er tale om en sæsoneffekt.

Det vil sige:

H_0: $\mu_{k.1} = \mu_{k.2} = \mu_{k.3} = \mu_{k.4}$ (ingen sæsoneffekt)
H_1: Mindst én middelværdi er forskellig (sæsoneffekt)

Nu anvendes enten lommeregneren eller Excel til at udføre beregningerne. På **lommeregneren** indtastes tallene bilaget i 4 registre *L1, L2, L3* og *L4*. Nu udføres H: ANOVA(*L1,L2,L3,L4*) → ENTER. Der fremkommer en udskrift, der kan bruges til at opstille ANOVA-tabellen. Bemærk at P-værdien også beregnes i lommeregnerens udskrift.

I **Excel** anvendes proceduren beskrevet i afsnit 4. Der fremkommer følgende udskrift:

Anova: Enkelt faktor

RESUME

Grupper	*Antal*	*Sum*	*Gennemsnit*	*Varians*
1 Kvartal	13	11379	875.31	44484.56
2 Kvartal	13	15011	1154.69	47490.56
3 Kvartal	13	11600	892.31	39166.40
4 Kvartal	13	19814	1524.15	240713.81

ANOVA

Variationskilde	*SK*	*fg*	*MK*	*F*	*P-værdi*	*F krit*
Mellem grupper	3587750	3	1195916.77	12.86	0.00	2.80
Inden for grupper	4462264	48	92963.83			
I alt	8050014	51				

Udskriften består af to dele. Øverst er der en resuméstatistik. Denne er velegnet til beregning af konfidensintervaller for middelværdien. Udskriftens anden del er ANOVA-tabellen. Den fundne F-værdi er lig med 12.86. Den kritiske værdi er lig med 2.80, og bliver givet direkte af udskriften. Da 12.86 > 2.80 forkastes H_0. Der er således et sæsonmønster. Dette fremgår også af P-værdien, der er langt under 0.05. Den kritiske værdi kan også findes i **Statistics Tables** på side 12. Her fås at værdien er lig 2.79, da jeg er nødt til at runde op til 50 frihedsgrader for fg_2.

Sæsonmønstret kan nu uddybes ved at opstille 95 % konfidensintervaller for de fire middelværdier. Ved anvendelse af udtrykket for konfidensintervallet på side 6 kan følgende opstilles:

$$\overline{X}_{K.1} \pm t_{0.025(n-p)}\sqrt{\frac{MSE}{n_1}} \Rightarrow 875.31 \pm t_{0.025(48)}\sqrt{\frac{92963.83}{13}} \Rightarrow 875.31 \pm 2.01(84.56) \Rightarrow [705.34\,;1045.28]$$

$$\overline{X}_{K.2} \pm t_{0.025(n-p)}\sqrt{\frac{MSE}{n_2}} \Rightarrow 1154.69 \pm t_{0.025(48)}\sqrt{\frac{92963.83}{13}} \Rightarrow 1154.69 \pm 2.01(84.56) \Rightarrow [984.39\,;1324.66]$$

$$\overline{X}_{K.3} \pm t_{0.025(n-p)}\sqrt{\frac{MSE}{n_3}} \Rightarrow 892.31 \pm t_{0.025(48)}\sqrt{\frac{92963.83}{13}} \Rightarrow 892.31 \pm 2.01(84.56) \Rightarrow [722.34\,;1062.28]$$

$$\overline{X}_{K.4} \pm t_{0.025(n-p)}\sqrt{\frac{MSE}{n_4}} \Rightarrow 875.31 \pm t_{0.025(48)}\sqrt{\frac{92963.83}{13}} \Rightarrow 1524.15 \pm 2.01(84.56) \Rightarrow [1354.18\,;1694.12]$$

Det ses, at der ikke er et overlap mellem fjerde kvartal og de øvrige tre kvartaler. Det vil sige, at mængden af reklamer er signifikant større sidst på året. Derimod er mængden af reklamer i andet kvartal ikke signifikant forskellig fra første og tredje kvartal.

Manuel beregning af F-testeren i ANOVA-tabellen

Er man ikke så heldig at have Excel eller den rette lommeregner til disposition, så må man ty til den manuelle beregning ☹. Til denne skal man have informationer om:

- Gennemsnittet for hver gruppe $\bar{x}_i$
- Antallet af observationer for hver gruppe n_i
- Variansen for hver gruppe s_i^2

Antag nu, at man er i besiddelse af disse informationer eksempelvis i form af resumé udskriften fra Excel. Det vil sige den øverste del af ANOVA-udskriften.

For at beregne F-testet skal man beregne elementerne i ANOVA-tabellen på side 5 nederst. Indledningsvis beregnes middelværdien for alle kaldet $\bar{x}$. Denne findes som et vægtet gennemsnit af middelværdien i grupperne:

$$\bar{x} = \frac{\sum_{i=1}^{p} n_i \bar{x}_i}{n} = \frac{13 \times 875.31 + 13 \times 1154.69 + 13 \times 892.31 + 13 \times 1524.15}{52} = \frac{57803.98}{52} = 1111.62$$

Nu beregnes variationen mellem grupperne, der er givet som $SSTR = \sum_{i=1}^{p} n_i(\bar{x}_i - \bar{x})^2$

$$\begin{aligned} SSTR = \sum_{i=1}^{p} n_i(\bar{x}_i - \bar{x})^2 &= 13 \times (875.31 - 1111.62)^2 + 13 \times (1154.69 - 1111.62)^2 \\ &+ 13 \times (892.31 - 1111.62)^2 + 13 \times (1524.15 - 1111.62)^2 \\ &= 725951.41 + 24115.32 + 625259.39 + 2212353.01 \\ &= 3587679.13 \end{aligned}$$

(Det er lettest at sætte 13 udenfor en parentes). Det er næsten den samme værdi, som i ANOVA-tabellen. Forskellen skyldes afrunding til 2 decimaler.

Variationen indenfor grupperne er givet som $SSE = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$. Se godt på dette udtryk. I hver gruppe skal man beregne kvadratsummen af observation til middelværdien. Det gør man også, når variansen beregnes. Her dividerer man også med (n_i−1). Det vil sige, at vi får, hvad der skal bruges ved at gange variansen med (n_i−1). For hver gruppe gælder der således, at $SSE_i = \sum_{i=1}^{n_i} (x_i = \bar{x}_i)^2 = (n_i - 1) \times s_i^2$. Da både observationer og variansen er kendt i alle grupperne kan *SSE* beregnes. Man får:

$$\begin{aligned} SSE = \sum_{i=1}^{p} (n_i - 1)s_i^2 &= (13-1) \times 44484.56 + (13-1) \times 47490.56 + (13-1) \times 39166.40 \\ &+ (13-1) \times 240713.81 \\ &= 533814.72 + 569886.72 + 469996.80 + 2888565.72 \\ &= 4462263.96 \end{aligned}$$

(Det er lettest at sætte (13−1) udenfor en parentes). Det passer også med den værdi, som findes i ANOVA-tabellen.

Afslutningsvis skal man beregne F-værdien. Ved anvendelse af opstillingen fra tabellen gøres dette som:

$$F = \frac{MSR}{MSE} = \frac{SSTR/(p-1)}{SSE/(n-p)} = \frac{3587679.13/(4-1)}{4462263.96/(52-4)} = \frac{1195893.04}{92963.83} = 12.86$$

Det var også den F-værdi, der blev opnået i ANOVA-tabellen.

Bilag: Vægt af reklametryksager indsamlet i 2004

Vægt af tryksager per uger i gram

Uge nr:	1 kv	2 kv	3 kv	4 kv
1	854	996	1,135	1,960
2	670	1,124	647	1,436
3	709	965	679	1,410
4	723	960	900	824
5	874	1,640	718	1,151
6	863	1,480	1,048	1,599
7	577	1,075	510	1,139
8	1,126	1,352	923	1,691
9	1,070	1,162	903	1,640
10	675	983	956	2,557
11	1,130	1,009	1,098	2,158
12	852	1,258	1,121	1,305
13	1,256	1,007	962	944

Nils Karl Sørensen ©
Statistik II, BA int, Flensborg

Sæt 4: Test af sammenhænge og fordelinger (χ^2−test)

af Nils Karl Sørensen

Indhold *side*

1. Goodness of Fit Test

Ofte kan man arbejde med datasæt, hvor man ikke kender den underliggende fordeling. Eksempelvis kan det være rart at vide, om et materiale følger en binomialfordeling, en normalfordeling eller en anden form for reference. Det kan eksempelvis være fordelingen af data i en totalpopulation, og så ønsker man at undersøge, om der er overensstemmelse med fordelingen i en stikprøve. Et test til at undersøge denne problemstilling benævnes et test for *Goodness of fit.* Testet anvendes således til at undersøge, om en fordeling er korrekt i forhold til data.

I testet betragtes et datasæt med frekvenser for et begrænset antal udfald. Det siges også, at testet er *enumerativt.* Testet er således perfekt til eksempelvis analyse af spørgeskemaer, hvor der måske er 5 eller 7 kategorier. Data antages at være opdelt i k grupper eller kategorier C, som det fremgår af tabellen[7].

Variabel A	C_1	...	C_i	...	C_k	Total
Frekvens	O_1	...	O_i	...	O_k	n

For hver kategori C_i er der en observeret værdi, der betegnes O_i. Summeres antallet af observerede værdier eller frekvenser, fås totalen i datasættet kaldet n. Ideen i testet er da at undersøge, om den observerende fordeling af data, er i overensstemmelse med et forventet datasæt E. Forventningen for et givet udfald i beregnes da som $E_i = np_i$. Sandsynlighederne p kan eksempelvis fås fra tabeller over binomial- eller normalfordelingen fra **Statistics Tables.**

[7] En sådan tabel med udfald fra 1 til 5 så vi også på i notesæt 5 til Statistik I om *Skalaniveauer og krydstabeller*. Der vendes tilbage til dette sæt af noter i det næste afsnit.

Med udgangspunkt i disse overvejelser formuleres hypoteserne som følger:

H_0: Datasættet kan beskrives ved anvendelse af den forventede reference
H_1: Datasættet kan *ikke* beskrives ved anvendelse af den forventede reference

Hvis H_0 accepteres, skal der være en så stor overensstemmelse mellem de observerede og forventede værdier, at afvigelsen mellem dem ikke er signifikant. Den forventede værdi er den, som er gældende under H_0. Denne er jo kendt fra teorien eller lignende, mens det jo for det observerede værdier netop undersøges, om disse data falder ind under den forventede eller teoretiske ramme.

Forskellen mellem E_i og O_i skal således være mindst mulig. Denne forskel kan være såvel positiv som negativ. For at løse dette problem kvadreres forskellen og summeres for alle kategorier. Der normaliseres med de forventede værdier, da disse er den kendte reference.

Forskellen mellem de forventede og de observerede værdier er et kvadrat. Derfor må den fordeling, som testet følger, også være et kvadrat. Testet må da være chi-i-anden fordelt.

Antallet af frihedsgrader for testet må være lig med de k kategorier fratrukket én frihedsgrad. Der mistes en frihedsgrad, da data antages at hidrøre fra en stikprøve.

Testeren er da

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Med frihedsgrader lig med $fg = k-1$. Yderligere skal der gælde, at $E_i \geq 5$

Eksempel på test af fordeling af karakterer
I Danmark er der en karakterskala, som afviger markant fra den tyske. I Tyskland anvendes der en skala med i alt 10 punkter for de beståede karakterer, mens skalaen i Danmark har 5 punkter for de beståede karakterer.

Den danske karakterskala er centreret omkring middelkarakteren 7. Så er der tre spring op til det næste niveau, der er 10. Endelig er der to spring op til topkarakteren 12, der gives for den maksimale målopfyldelse af et bestemt fags fagbeskrivelse. Tilsvarende den anden vej, hvor karakteren under 7 er tre spring ned til 4. Dernæst er der to spring til karakteren 2, der netop sikrer beståelse[8]. Karakterskalaen fremgår af den første linje i tabellen nedenfor.

[8] Karakterskalaen er udførligt omtalt i: Undervisningsministeriet (2004): **Betænkning om indførelse af en ny karakterskala til erstatning af 13-skalaen afgivet af karakterkommissionen, november 2004.** Betænkning nr. 1453. Det er især kapitel 8, hvor den nye karakterskala udvikles, der er interessant. Forfatteren til disse noter har været i dialog med udvalgets formand, som har bekræftet, at det er den skitserede tankegang, der ligger bag.

Hvis karakterskalaen er centreret omkring middelkarakteren 7 kan følgende spørgsmål stilles: Hvor mange studenter skal have 7, og hvordan skal fordelingen være af de øvrige karakterer?

Til eksamen er kravet normalt, at 50 % af svarene skal være korrekte for at sikre bestået. Hvordan skal dette fortolkes i forhold til den danske karakterskala? Ved eksamen er der to udfald; nemlig bestået eller ikke bestået. De to udfald er lige sandsynlige. Det vil sige, at den underliggende karakterfordeling for de bestået karakterer følger en Binomialfordeling med $p = 0.5$ og 5 udfald. En stokastisk variabel med 5 udfald har udfaldene *0,1,2,3 og 4*. Det vil sige $n = 4$. Dette kan umiddelbart virke mærkeligt, men husk på at vi kun ser på de *bestået karakterer*. Der er således 5 *udfaldsrum*. Sandsynlighedsfordelingen for en binomialfordelt stokastisk variabel med 5 udfald og $p = 0.5$ findes i **Statistics Tables** på side 2. Disse sandsynligheder fremgår af den tredje linje af tabellen nedenfor.

Lad os nu undersøge en specifik problemstilling. I tabellen nedenfor er i anden linje angivet fordelingen af danske karakterer for de studenter, der bestod international økonomi (VWL-III) ved Flensburg Universitet ved eksamen på BA-int studiet februar 2011. Dette materiale blev også anvendt i notesæt 1 om *Deskriptiv Statistik* i Statistik I

Hypoteserne for testet er nu:

H_0: Fordelingen af karakterne ved VWL-III følger binomialfordelingen
H_1: Fordelingen af karakterne ved VWL-III følger *ikke* binomialfordelingen

Med udgangspunkt i den fundne total på 92 og sandsynlighederne fundet for den givne binomialfordeling, kan de forventede værdier beregnes. De forventede værdier findes i tabellens fjerde linje. I den femte linje beregnes testeren. Det gule felt yderst til højre markerer chi-i-anden værdien, som er fundet ved at summere vandret i tabellen.

Karakter	2	4	7	10	12	Total
Observeret (O_i)	10	26	33	19	4	92
Sandsynlighed (p_i)	0.0625	0.250	0.375	0.250	0.0625	1.000
Forventet ($E_i = np_i$)	5.75	23	34.5	23	5.75	92
$(O_i-E_i)^2/E_i$	3.141	0.391	0.065	0.696	0.533	4.826

Testeren findes at være lig med $\chi^2 = \sum_{i=1}^{c} \frac{(O_i - E_i)^2}{E_i} = 4.826 \approx 4.83$

Antallet af frihedsgrader er lig med $fg = k-1 = 5-1 = 4$.

Den kritiske værdi findes ved anvendelse af **Statistics Tabels** side 11 til at være lig med $\chi^2_{(4)0.05} = 9.487$ da $\alpha = 0.05$. Da $4.826 < 9.487$ accepteres H_0. Det vil sige, at fordelingen af karaktererne følger en binomialfordeling med de skitserede parametre.

Goodness of fit testet kan beregnes ved anvendelse af **lommeregneren.** Anvend STAT → TESTS → D: GOF-Test → ENTER

Data tastes ind i registrene L_1 og L_2. Det vil sige, at man selv skal beregne de forventede værdier. Endvidere angives antallet af frihedsgrader. Dernæst anvendes CALCULATE → ENTER.

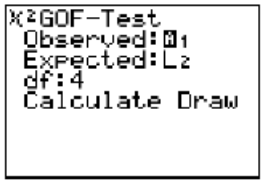
Udskriften giver værdien for chi-i-anden samt *p-værdien*, der i eksemplet ovenfor er lig med 0.3056. Da *p-værdien* er større end 0.05 passer det fint med resultatet ovenfor.

Anvender man i stedet DRAW, så får man en fin tegning af chi-i-anden fordelingen med arealet for *p-værdien* skraveret (dette forudsætter at lommeregnernes figurfunktion er korrekt indstillet).

Se også følgende eksempel som illustration fra manualen til lommeregneren:

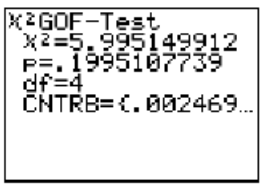
I eksemplet:
list 1={16,25,22,8,10}
list 2={16.2,21.6,16.2,14.4,12.6}

Inputskærmbilledet Chi-square Goodness of Fit:

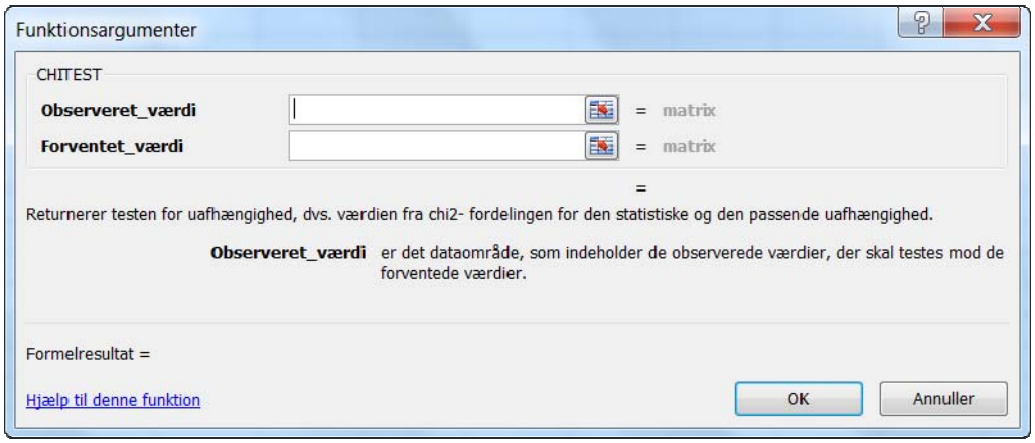


Bemærk: Tryk på STAT ▸ ▸ for at vælge **TESTS**. Tryk flere gange på ▾ for at vælge **D:X²GOF-Test...** Tryk på ENTER. Tryk på ▾ ▾ ▾ for at indtaste data for df frihedsgrad (degree of freedom). Tast 4.

Beregnede resultater:



Ved anvendelse af **Excel** tastes **Formler / Indsæt funktion / statistisk / Chitest**, hvorefter der fremkommer:



Bemærk, at såvel de observerede værdier som de forventede værdier også her skal beregnes på forhånd. Beregningen giver kun p-værdien for udfaldet af testet.

2. Test for uafhængighed

I det forrige afsnit, så vi på, hvordan man for enkelt variabel kan undersøge, om et datasættet følger en given fordeling eller reference.

Denne problemstilling kan udvides til at omfatte to variabler. I sæt 5 i noterne til Statistik I fra efteråret definerede vi en tabel, som kombinerede to elementer i den samme tabel. Sådan en tabel kaldes en *krydstabel* eller en *antalstabel*[9]. I Statistik I arbejdede vi med at opstille tabeller, der viste en sammenhæng. I dette afsnit anvendes nu den viden, som vi i den mellemliggende tid har lært om hypoteser og fordelinger, til at opstille et specifikt test for *uafhængighed*.

Et test for uafhængighed mellem to variabler er netop et test for, om to variabler i en krydstabel er relateret til hinanden. Er der *afhængighed* mellem to variabler, betyder det, at de er relateret til hinanden. Det kan eksempelvis tænkes at være tilfældet mellem rygning og risikoen for cancer eller mellem antallet af studenter, der har fulgt undervisningen i et fag, og antallet af studenter der består faget.

Ordet *uafhængighed* har vi allerede stiftet bekendtskab med i det sæt af noter fra Stsitik I, der omhandlede sandsynlighedsteori[10]. Var der uafhængighed mellem to variabler kaldet A og B, så kunne disse multipliceres. Var der uafhængighed, var der specielt gældende at $P(A \cap B) = P(A)P(B) = 0$.

[9] Sæt 5: Skalaniveauer og krydstabeller
[10] Sæt 2: Sandsynlighedsteori og statistiske fordelinger

Var der imidlertid *afhængighed* mellem to variabler, så var der en fællesmængde $P(A \cap B)$ af udfald, som ikke var lig med nul. Det vil sige, at $P(A \cap B) = P(A)P(B) \neq 0$ Denne fællesmængde skulle man huske ikke at tælle dobbelt, når man skulle beregne sandsynligheder. I relation til vort test implicerer denne situation netop, at der er afhængighed og at variablerne er relateret til hinanden. Denne situation er som regel den mest interessante.

Dette fohold kan anvendes til at opstille et test efter de samme retningslinier, som vi netop har gjort det i det forrige afsnit. Det vil sige et test, der er baseret på χ^2 eller chi-i-anden fordelingen, samt på beregninger af forventede værdier.

Indledningsvis opstilles hypoteserne for testet som:

H_0: De to variabler er uafhængige $P(A \cap B) = P(A)P(B)$
H_1: De to variabler er afhængige $P(A \cap B) \neq P(A)P(B)$

Materialet kan være opstillet i en *krydstabel* skitseret som følger:

Opstilling af krydstabel

	Variabel A					
Variabel B	**1**	**2**	**3**	**4**	**5**	**Total**
1	O_{11}	O_{12}	O_{13}	O_{14}	O_{15}	R_1
2	O_{21}	O_{22}	O_{23}	O_{24}	O_{25}	R_2
3	O_{31}	O_{32}	O_{33}	O_{34}	O_{35}	R_3
4	O_{41}	O_{42}	O_{43}	O_{44}	O_{45}	R_4
Total	C_1	C_2	C_3	C_4	C_5	n

Her benævnes en given observation som O_{ij}. I krydstabellen er antallet af rækker R_i, $i=1,2,...,r$ mens antallet af kolonner C_j, $j=1,2,...,c$. I eksemplet i tabellen er $r = 4$ og $c = 5$. Antallet af observationer er lig med n.

Testeren kan findes som:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Antallet af frihedsgrader er lig med $fg = (r-1)(c-1)$

Det fremgår, at testet er meget lig testet for goodness of fit. Nu haves der to summer, og som det vil fremgår, beregnes de forventede værdier lidt anderledes.

Den forventede værdi for et givet element i række i og kolonne j benævnes E_{ij}. For at kunne beregne de forventede værdier anvendes forudsætningen om uafhængighed, som netop skal

være gældende under H_0. I konteksten af en krydstabel med dimensionen $r \times c$, er gældende at den forventning, der er associeret med celle *(i,j),* er lig med $E_{ij} = nP(i \cap j)$. Her gælder at $P(i \cap j) = P(i)P(j)$ som en følge af antagelse om uafhængighed.

Fra summen af elementerne i en række kan forventningen for hændelsen *i* findes som R_i/n. Tilsvarende kan man fra summen af elementerne i en kolonne finde forventningen for hændelsen *j* findes som C_j/n.

Ved substitution af de marginale forventninger kan den forventede værdi for en givet celle findes som $(i, j) : E_{ij} = nP(i \cap j) = n(R_i / n)(C_j / n) = R_i C_j / n$.

Forventningen er således $E_{ij} = \frac{R_i C_j}{n}$

Specialtilfælde
Ved en undersøgelse af en krydstabel med kun to udfald for hver variabel, kan der opstå en speciel situation. Ved en 2×2 krydstabel haves nemlig kun én frihedsgrad. I dette tilfælde er der en tendens til at testeren overvurderes. Derfor anvendes *Yates korrektionen.* Her fratrækkes 0.5 fra den numeriske forskel mellem den observerede og den forventede værdi. Testeren er da lig med

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(|O_{ij} - E_{ij}| - 0.5)^2}{E_{ij}}$$

Med frihedsgrader lig med 1.

Styrken af testet for uafhængighed
Styrken af testet for uafhængighed kan undersøges ved *Cramer's V*. Dette mål er som en korrelation, som er kendt fra noterne fra statistik I. Korrelationen vil variere mellem 0 og 1. Udtrykket er gengivet i Müller-Benedicts bog, men findes ikke gennemgået i ret mange bøger om statistik (denne forfatter har ikke kunne finde den i andre bøger i mine omkring 30 år som underviser). *Cramer's V* er givet som:

$$V = \sqrt{\frac{\chi^2}{n \times \min(r-1; c-1)}}$$

Her angiver min(r−1;c−1), at man skal anvende den *mindste* af enten rækkerne eller kolonnerne fratrukket 1.

Eksempel
I notesættet om *skalaniveauer og krydstabeller* opstillede vi en tabel for valg af cola mellem drenge og piger. Ved en spørgeskemaundersøgelse blev 77 studerende ved SDU Campus Sønderborg spurgt om deres foretrukne type af coca-cola. Der kunne svares mellem almindelig coca-cola og diæt coca-cola. Ved undersøgelsen angav studenterne tillige deres køn. Svarende blev optalt og samlet i en *krydstabel,* idet totalerne også blev angivet. Denne tabel er vist nedenfor til venstre.

I notesættet om *skalaniveauer og krydstabeller* fandtes det intuitativt, at der var *afhængighed* mellem valg af cola og køn. Kvinderne foretrak diæt colaen, mens mændene foretrak den almindelig cola. Dernæst blev der opstillet en tabel, hvor der blev postuleret, at der ikke var afhængighed mellem de to variabler.

Nu er vi i en position, så der kan opstilles et mere formelt test for, om der er afhængighed mellem køn og valg af cola. Indledningsvis opstilles hypoteserne:

H_0: Der er ingen sammenhæng mellem køn og valg af cola (uafhængighed)
H_1: Der er sammenhæng mellem køn og valg af cola (afhængighed)

Først beregnes de forventede værdier ved anvendelse af række- og kolonnesummerne. For eksempel findes for kvinders valg af cola:

$$E_{11} = \frac{36 \times 33}{77} = 15.43 .$$

Tilsvarende for de øvrige forventede værdier. Det er let at se, om man har regnet rigtigt, da der ikke ændres på række- og kolonnesummerne. De forventede værdier er vist i tabellen nedenfor til venstre.

Observeret			
	Cola	Diæt cola	Total
Kvinde	3	33	36
Mand	30	11	41
Total	33	44	77

Forventet			
	Cola	Diæt cola	Total
Kvinde	15.43	20.57	36
Mand	17.57	23.43	41
Total	33	44	77

Nu kan testeren beregnes:

$$\chi^2 = \frac{(3-15.43)^2}{15.43} + \frac{(33-20.57)^2}{20.57} + \frac{(30-17.57)^2}{17.57} + \frac{(11-23.43)^2}{23.43} = 10.01 + 7.51 + 8.79 + 6.59 = 32.90$$

Antallet af frihedsgrader er lig med $fg = (c-1)(r-1) = (2-1)(2-1) = 1$

Antages et signifikansniveau på 95 procent svarende til $\alpha = 0.05$, så kan den kritiske værdi findes i **Statistics Tables** side 11 til at være lig med $\chi^2_{(1)} = 3.84$. Da 32.90 > 3.84 forkastes H_0. Der findes således en sammenhæng mellem køn og valg af cola. Kvinderne foretrækker diæt cola, mens mændene foretrækker almindelig cola.

Styrken af undersøgelsen kan findes ved at beregne *Cramer's V*. Her fås, da *r = c = 2.*

$$V = \sqrt{\frac{\chi^2}{n \times \min(r-1; c-1)}} = \sqrt{\frac{32.90}{77 \times (2-1)}} = 0.65$$

Det vil sige en ganske stærk sammenhæng.

Da antallet af frihedsgrader er lig med ét, så kan man beregne Yates korrektionen. Den er lig med:

$$\chi^2 = \frac{(|3-15.43|-0.5)^2}{15.43} + \frac{(|33-20.57|-0.5)^2}{20.57} + \frac{(|30-17.57|-0.5)^2}{17.57} + \frac{(|11-23.43|-0.5)^2}{23.43}$$
$$= 9.22 + 6.92 + 8.10 + 6.07 = 30.31$$

Som det ses falder værdien af testeren. Det vil sige, at sandsynligheden for at acceptere H_0 bliver reduceret. Da 30.31 > 3.81 fastholdes den tidligere konklusion.

Test for uafhængighed kan også udføres på **lommeregneren.** Anvend STAT → TESTS → C: χ^2 -Test → ENTER

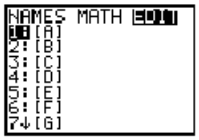
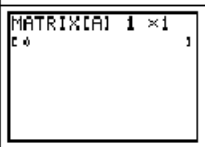
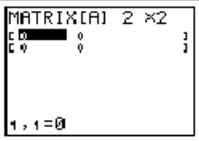
Nu skal såvel de observerede som de forventede data angives i matrixform som henholdsvis [A] og [B]. Det vil sige, at de forventede værdier skal beregnes på forhånd. Dernæst anvendes CALCULATE → ENTER.

Udskriften giver værdien for chi-i-anden samt *p-værdien*. Antallet af frihedsgrader beregnes automatisk.

Hvordan opstilles data i matrixform? Test 2ND → MATRIX. Nu kommer en menu, hvor man kan definere matrixer. Vælg EDIT og 1: [A] → ENTER. Da fremkommer en menu, hvor først matrix' dimension vælges. Dernæst indtastes de observerede værdier. Tilsvarende gøres for de forventede værdier i matrixen [B].

Se nedenstående eksempel fra instruktionsbogen til TI-84'eren.

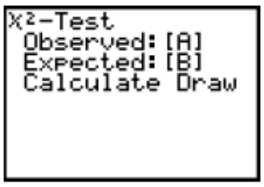
Oprettelse af en ny matrix

Tryk på	Resultat
2nd [MATRIX] ▪	
ENTER	
2 ENTER **2** ENTER	
1 ENTER **5** ENTER **2** ENTER **8** ENTER	

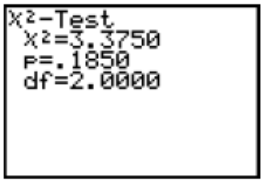
Bemærk: Når du trykker på ENTER, flytter markøren automatisk til næste celle og fremhæver den, så du kan fortsætte med at indtaste eller redigere værdier. Hvis du vil indtaste en ny værdi, kan du begynde uden at trykke på ENTER, men du skal anvende ENTER for at redigere en eksisterende værdi.

Et eksempel på udførsel af et test er som følger:

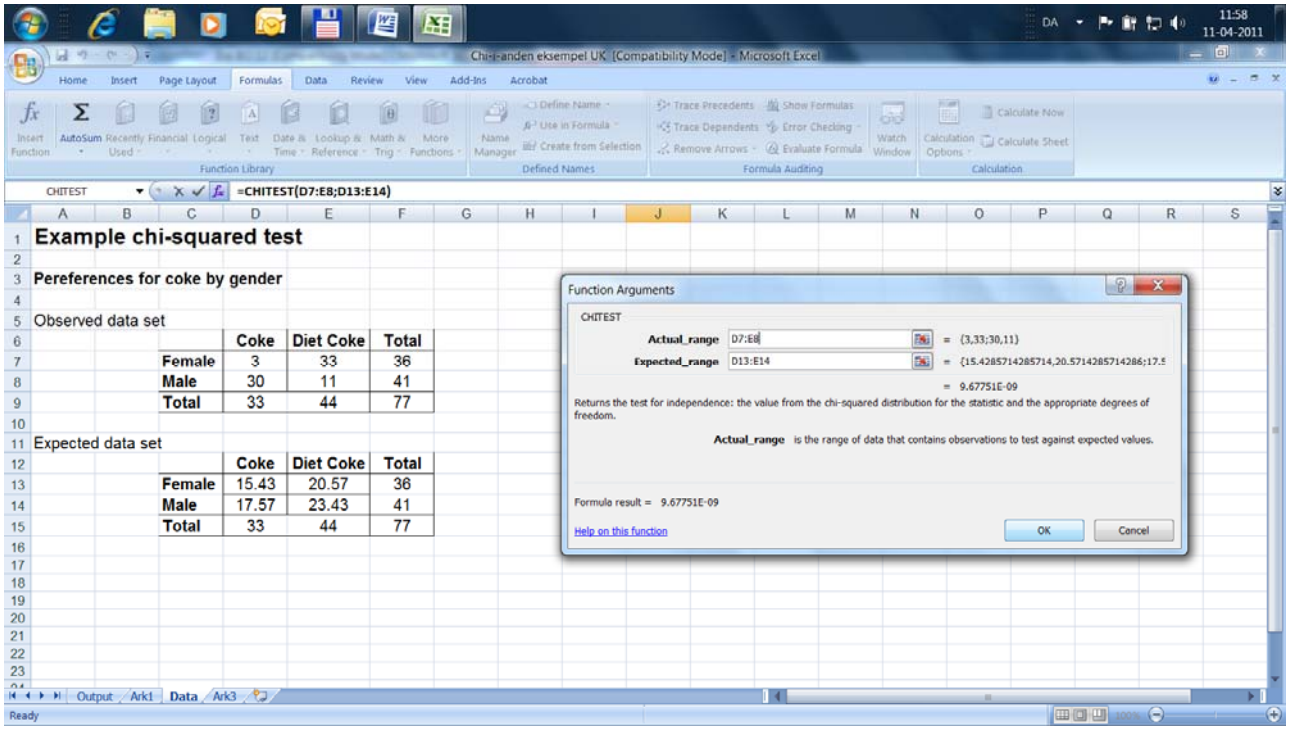
Input:



Beregnede resultater:



Ved anvendelse af **Excel** gås der frem som under goodness of fit testet. Det vil sige, at der tastes **Formler / Indsæt funktion / statistisk / Chitest**. Såvel de observerede som de forventede værdier skal beregnes på forhånd. Bemærk at række- og kolonnetotalerne *ikke* skal markeres. I eksemplet fås følgende skærmbillede:


Chi-i-anden eksempel UK [Compatibility Mode] - Microsoft Excel
=CHITEST(D7:E8;D13:E14)
Example chi-squared test
Pereferences for coke by gender
Observed data set
Coke Diet Coke Total
Female 3 33 36
Male 30 11 41
Total 33 44 77
Expected data set
Coke Diet Coke Total
Female 15.43 20.57 36
Male 17.57 23.43 41
Total 33 44 77
Function Arguments
CHITEST
Actual_range D7:E8 = {3,33;30,11}
Expected_range D13:E14 = {15.4285714285714,20.5714285714286;17.5
= 9.67751E-09
Returns the test for independence: the value from the chi-squared distribution for the statistic and the appropriate degrees of freedom.
Actual_range is the range of data that contains observations to test against expected values.
Formula result = 9.67751E-09
Help on this function
OK
Cancel
Output Ark1 Data Ark3
Ready

Nils Karl Sørensen ©
Statistik II, BA int, Flensborg

Sæt 5: Regressionsanalyse

af Nils Karl Sørensen

Indhold *side*

1. Simpel regression

Regression er en af de mest hyppigt anvendte metoder til at analysere sammenhænge mellem stokastiske variable. I *regressionsanalyse* forudsiger man udfaldet af en stokastisk variabel y ved anvendelse af en eller flere stokastiske variabler x. I regressionsanalysen anvendes mange af de ting, som vi har beskæftiget os med i de foregående sæt af noter som eksempelvis korrelation, t-test og p-værdier. Regressionsanalysen falder i naturlig forlængelse af den deskriptive statistik. Her blev sammenhænge mellem datasæt beskrevet, men ikke sat ind i en statistisk teoretisk konsistent referenceramme. Dette gøres ved regression.

Ordet *regression* er afledt af det engelske *regress* på dansk: ”at vende tilbage” eller ”tilbagevenden”. Baggrunden for anvendelsen af ordet er, at Charles Darwins fætter, Francis Galton i 1889 udgav en bog om analyser af arvelighed. I Galtons bog var der et korrelationsdiagram (XY-diagram), der undersøgte sammenhængen mellem højden af fædre og sønner. Fædrenes højde blev målt på den vandrette x-akse, mens sønnernes højde blev målt på den lodrette y-akse. Galton kunne finde disse data fra målinger af værnepligtige soldater, hvor man både havde informationer om sønner og fædre omkring deres attende år. Galton fandt til sin overraskelse, at den rette linje, der passede bedst til punkterne i diagrammet, havde en hældningskoefficient, der var mindre end ét. Høje fædre får sønner, der generelt er mindre end dem selv, mens fædre med en højde under gennemsnittet generelt får sønner, der er højere end dem selv. Der foregår således et tilbagefald (regress) af højden mod midten.

Galtons undersøgelse byggede på en fejlslutning (Galtons fallacy). Dette skyldes, at de to målinger af fædre og sønner var adskilt i tid. Den naturlige variation i undergrupper fra en population vil bevirke, at nogle fædre uden genetiske anlæg for ”højhed” tilfældigvis er

høje. Det er overvejende sandsynligt, at disse fædre vil få sønner, der er mindre end dem selv. Omvendt er det mest sandsynligt, at mindre høje fædre uden genetisk anlæg for "lavhed", vil få sønner, der er højere end dem selv. Det er dette forhold, der implicerer Galtons "tilbagefald" mod midten

Den simple model og dennes forudsætninger
Som udgangspunkt for den simple regression betragtes en variabel *y*, som søges forklaret lineært af en anden variabel *x*. I økonomi betegnes variablen *y* ofte som den endogene/forklarede eller *afhængige* variabel, mens *x* er den eksogene/forklarende eller *uafhængige* variabel.

Den lineære sammenhæng beskrives da som:

$$y_i = E(y_i) + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \text{hvor } i = 1,2,\ldots,n$$

Her er β_0 konstantleddet, mens β_1 angiver hældningen på linjen. Denne kan være såvel positiv som negativ. Variablen *y* skal forklares af variablen *x*. Fodtegnet *i* angiver antallet af par af observationer af *x* og *y*. Der er *n* par af observationer. Observationerne kan enten være angivet over tid eller som opgjort på samme tid. Førstnævnte kaldes et *tidsserie* datasæt, mens sidstnævnte er et *tværsnit* datasæt.

Parametrene β_0 og β_1 skal fastlægges ud fra data ved anvendelse af matematiske metoder. Når dette er foretaget, kan man beregne eller *prædikte* værdien af *y* kaldet *E(y)*. Det er ikke sikkert, at man rammer korrekt. Forskellen mellem den prædiktede værdi af *y* kaldet $\hat{y}$ og det observerede *y* kaldet *residualet* og delfineres $\varepsilon = \hat{y} - y$.

Nu ønskes det at fastlægge værdierne for parametrene, således at der fremkommer den bedst mulige linje. For at kunne gøre dette, må der opstilles et kriterium for beregningen. Umiddelbart er det rimeligt at ønske, at den opstillede model er bedst mulig til at predikte *y*. Forventningen til *y* skal således have en så høj grad af overensstemmelse med den observerede værdi som muligt. Et udtryk for denne forskel er residualet. Parametrene skal således fastlægges *så residualerne minimeres.*

Ud fra dette kriterium kan der opstilles følgende forudsætninger for den simple regressionsanalyse:

- Relationen mellem *x* og *y* skal være lineær
- Residualerne skal være normalfordelte med middelværdi nul og konstant varians. Dette skrives som $\varepsilon \approx NID(0, \sigma_\varepsilon^2)$
- Residualerne skal være uafhængige. Det vil sige, at et residual ε skal være uafhængig af ethvert andet residual og af dermed af *y*

Opstilling og løsning af den simple model

For et givet sæt af observationer (x_i,y_i) kan den *simple regressionsmodel* opstilles som følger:

$$y_i = E(y_i) + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \text{hvor } i = 1,2,\ldots,n$$

Hvor n betegner antallet af observationer.

Det ønskes nu at fastlægge parametrene β_0 og β_1 således at residualerne minimeres. Disse er defineret som $\varepsilon = \hat{y} - y$. Det vil sige, at et residual kan være såvel positivt som negativt. På denne måde kan positive og negative residualerne gå ud mod hinanden ganske som det er tilfældet, når man eksempelvis beregner variansen.

For at kunne løse dette problem minimeres i stedet kvadratet af residualet ε_i^2. På denne måde fastholdes al variation i materialet. Da man betragter observationer fra $i = 1,2, \ldots n$ minimeres *summen af kvadraternes afvigelser.* Teknisk opskrives minimeringsproblemet for de to parametre da som

$$\text{Minimer } \varepsilon^2 \text{ med hensyn til } \beta_0 \text{ og } \beta_1 \text{ eller } \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Dette system kan løses som et almindeligt minimeringsproblem, hvor man finder de partielle afledede for de to parametre β_0 og β_1. Der fremkommer et system af to ligninger med to ubekendte. Løsningerne til systemet benævnes *normalligningerne.*

Mens metoden til løsningen af systemet er kompliceret, så er løsningerne ligetil at anvende. Definer følgende kvadratsummer for x, y og xy:

$$SS_x = \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad \text{hvor } \bar{x} \text{ er middelværdien af } x$$

$$SS_y = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad \text{hvor } \bar{y} \text{ er middelværdien af } y$$

$$SS_{xy} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})$$

Løsningerne til parametrene i regressionen er da givet som:

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \text{ og } \beta_1 = \frac{SS_{xy}}{SS_x}$$

Løsningerne, der fremkommer ved denne procedure, kaldes mindste kvadraters metode (På engelsk: *OLS* or ***O**rdinary **L**east **S**quares*[11]). Ved anvendelse af denne metode gælder specielt:

> *Gauss-Markow Theorem:*
> Mindste kvadraters metode (OLS) giver den bedste lineære estimation (BLUE) af β'erne. (Engelsk: The Best Linear Unbiased Estimate (BLUE))

Dette er i et vigtigt resultat i en statistisk kontekst. Under de givne forudsætninger vil anvendelsen af normalligningerne give en korrekte værdi af parametrene. I *residualanalysen* undersøges om forudsætningerne for regressionsmodellen er opfyldt. Dette vendes der tilbage til senere.

Eksempel på beregning af parametrene β_0 og β_1

Betragt nu følgende eksempel til illustration. For variablerne y_i og x_i haves følgende oplysninger for i alt 7 par af observationer:

y_i	40	50	50	70	65	65	80
x_i	100	200	300	400	500	600	700

Umiddelbart ses det, at når x stiger i værdi, så stiger y. Det vil sige, at der må forventes en positiv sammenhæng. Der opstilles en tabel til at finde mellemregninger og værdierne for β_0 og β_1

y_i	x_i	$(x_i - \bar{x})$	$SS_x = (x_i - \bar{x})^2$	$(y_i - \bar{y})$	$SS_y = (y_i - \bar{y})^2$	$SS_{xy} = (x_i - \bar{x})(y_i - \bar{y})$
40	100	−300	90,000	−20	400	6,000
50	200	−200	40,000	−10	100	2,000
50	300	−100	10,000	−10	100	1,000
70	400	0	0	10	100	0
65	500	100	10,000	5	25	500
65	600	200	40,000	5	25	1,000
80	700	300	90,000	20	400	6,000
$\sum y_i$=420	$\sum x_i$ = 2,800		$\sum$ = 280,000		$\sum$ = 1,150	$\sum$ = 16,500

Middelværdierne er da: $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{420}{7} = 60$ og $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{2{,}800}{7} = 400$

Ved at indsætte i normalligningerne findes løsningerne for β_0 og β_1:

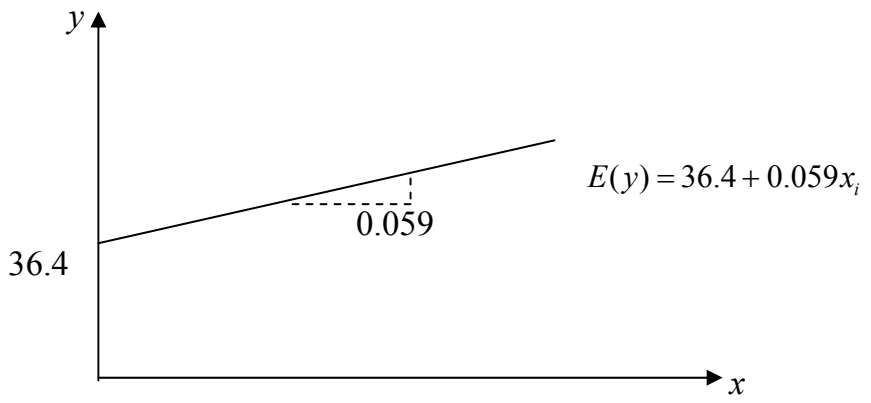
[11] Denne løsningsmetode blev udviklet Carl Friedrich Gauss (1777−1855), der også udviklede normalfordelingen omtalt i noterne til Statistik I. Han anvendte imidlertid matrix algebra. Matematikeren Thomas Jordan udviklede en meget elegant metode til direkte at løse minimeringsproblemet. Dette kaldes ofte Gauss-Jordan metoden.

$$\beta_1 = \frac{SS_{xy}}{SS_x} = \frac{16{,}500}{280{,}000} = 0.059$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} = 60 - 0.059(400) = 36.4$$

Regressionen er lig med: $\hat{y}_i = \beta_0 + \beta_1 x_i = 36.4 + 0.059 x_i$

Regressionslinjen er indtegnet i figuren. Det er en meget flad kurve med en positiv hældning.



Korrelation og determination

Vi har tidligere i Statistik I beregnet korrelationen mellem *x* og *y*. Korrelationen angiver graden af samvariation mellem de to variabler, der indgår i regressionen. Korrelationen, der varierer mellem $-1 \le r \le 1$, er defineret som:

$$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

For eksemplet ovenfor kan ***korrelationen*** beregnes til:

$$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} = \frac{16{,}500}{\sqrt{280{,}000(1{,}150)}} = 0.9195$$

Korrelationen er positiv som forventet.

I forlængelse af korrelationen kan man beregne ***determinationen*** også kaldet R^2. Den er defineret som:

$$R^2 = r^2 = (0.9195)^2 = 0.8455$$

Det vil sige, at determinationen er lig med kvadratet på korrelationen. Determinationen er således altid positiv. I dette tilfælde forklarer regressionslinjen 84.55 % af variationen i *y*.

Determinationskoefficienten udtrykker, hvor stor en del af variationen i materialet, som forklares af regressionslinjen og varierer $0 \le R^2 \le 1$. Desto højere værdi af R^2, desto bedre er regressionen.

Determinationskoefficienten R^2 kan også beregnes ved anvendelse af de beregnede værdier i tabellen ovenfor. Værdien SS_y udtrykker den samlede variation i *y*, mens $SS_y - \beta_1 SS_{xy}$ udtrykker den variation i modellen, som ikke forklares af regressionen. Kald sidstnævnte for *SSE* (**S**um **S**quared **E**rror) og førstnævnte for *SST* (**S**um **S**quared **T**otal), da kan R^2 beregnes som:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 = SS_y - \frac{SS_{xy}^2}{SS_x} = SS_y - \beta_1 SS_{xy}$$
$$= 1{,}150 - (0.059)16{,}500 = 177.68$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = SS_y = 1{,}150$$

Man finder nu R^2 som:

$$R^2 : 1 - \frac{SSE}{SST} = 1 - \frac{177.68}{1{,}150} = 0.8455 \qquad \text{(som forventet)!}$$

Determinationskoefficienten er særdeles udbredt som fortolkningsparameter for en regression. Målet skal imidlertid anvendes forsigtigt i modeller med mange *x*-variabler. Her kan modellen blive "overidentificeret", hvis flere af variablerne "forklarer hinanden". I sådanne tilfælde kan værdien af determinationskoefficienten falde. Derfor beregnes ofte *den justerede determinationskoefficient* også benævnt R^2-adj. Her korrigeres der for det netop skitserede forhold. R^2-adj er givet som:

$$R^2_{adj} : 1 - \frac{SSE/[n-(k+1)]}{SST/(n-1)}$$

Her er *k* antallet af *x*-variable i regressionen. I den simple regression gælder der altid at *k = 1*. Da det huskes, at der er 7 observationer i eksemplet fås:

$$R^2_{adj} : 1 - \frac{SSE/[n-(k+1)]}{SST/(n-1)} = 1 - \frac{177.68/[7-(1+1)]}{1150/(7-1)} = 1 - \frac{35.536}{191.667} = 0.8146$$

Bemærk at R^2-adj altid er mindre end R^2. Desto større datasæt, desto mindre bliver forskellen mellem de to mål.

Standardfejlen

Der er nu beregnet værdier for parametrene samt et udtryk for, hvor god regressionen er. Hvordan kan man beregne et udtryk for usikkerheden på hældningen? Til dette formål skal der beregnes et generelt usikkerhedsmål for hele regressionen.

Standardfejlen er et udtryk for modellens generelle usikkerhed eller variation. Den beregnes næsten som standardafvigelsen og betegnes da også med *s*. Den er givet som:

$$s = \sqrt{\frac{SSE}{n-(k+1)}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-(k+1)}} = \sqrt{MSE}$$

Hvor *MSE* betegner *middelkvadratfejlen* (**M**ean **S**quare **E**rror). I eksemplet kan *s* beregnes som:

$$s = \sqrt{\frac{SSE}{n-(k+1)}} = \sqrt{\frac{177.68}{7-(1+1)}} = 5.96$$

Standardfejlen i en regression skal være så lille som muligt.

Sammenlignes flere forskellige regressioner skal der således gælde:

- Korrelationen skal være så tæt på −1 eller 1 som muligt
- Determinationskoefficienten skal være så tæt på 1 som muligt
- Standardfejlen skal være så lille som muligt

Standardafvigelsen for β_1 (hældningen) og konfidensinterval for denne

Man kan finde standardafvigelsen for hældningskoefficienten ved anvendelse af følgende udtryk, hvor de tidligere fundne værdier er indsat:

$$s(\beta_1) = \frac{s}{\sqrt{SS_x}} = \frac{5.96}{\sqrt{280{,}000}} = 0.01126$$

Her anvendes foruden standardfejlen variationen på *x*, da det jo er denne variabel, der opstilles et usikkerhedsmål for. Med kendskab til standardafvigelsen kan der nu opstilles et konfidensinterval for hældningskoefficienten β_1. Et 95 procents konfidensinterval, det vil sige med $\alpha = 0.05$, er defineret som:

$$\left[\beta_1 \pm t_{\alpha/2} s(\beta_1)\right] \Rightarrow \left[0.059 \pm 2.571(0.01126)\right] \Rightarrow \left[0.059 \pm 0.0289\right] \Rightarrow \left[0.0299; 0.0879\right]$$

Ved et 95 procents signifikansniveau fås at α/2 = 0.025, da konfidensintervallet er tosidet. Konfidensintervallet er t-fordelt, da datasættet er lille. Antallet af frihedsgrader er lig med *fg = n − (k+1).* I eksemplet haves at *n = 7* og *k = 1.* Det betyder, at *fg = n − (k+1) = 7 – (1+1) = 5.* I **Statistics Tables** findes t-værdien til at være lig med 2.571. Dette anvendes i formlen ovenfor.

Ganske som det var tilfældet i statistik I, kan man anvende konfidensintervallet til at undersøge hypoteser. Det er mest almindeligt at undersøge, om β_1 er forskellig fra nul. Dette er det samme som at sige at variablen *x* har en mening for *y*. Hypoteserne er da lig med:

H_0: $\beta_i = 0$	variablen *x* har ikke betydning
H_1: $\beta_i \neq 0$	variablen *x* har betydning

Hvis værdien nul indgår i konfidensintervallet accepteres H_0, mens H_1 accepteres, hvis værdien nul ikke indgår i konfidensintervallet. Accepteres H_0 har regressionen ikke nogen relevans. Man kan også teste, om en bestemt værdi ligger inden for konfidensintervallet ved simpel sammenligning. I eksemplet kan hældningskoefficienten ikke være lig med 0.1, da denne værdi falder udenfor konfidensintervallet.

Testeren kan også opstilles mere formelt, som det blev gjort i noterne om tests. Her er testeren lig med, idet de relevante værdier er indsat:

$$t = \frac{\beta_1}{s(\beta_1)} = \frac{0.05892}{0.01126} = 5.2327$$

Frihedsgraderne er som ovenfor og t-værdien er lig med 2.571. Da 2.571 < 5.2327 accepteres H_1 som forventet, og *p-værdien* ligger langt under 0.05.

Prædiktion

En prædiktion er en *forudsigelse.* Der indsættes en værdi for *x* i den estimerede regression og på denne måde beregnes en *prædiktion* for *y.* Kald prædiktionen for x_0. Antag i vores eksempel, at $x_0 = 650$. Da kan den tilhørende værdi for *y* kaldet $\hat{y}$ beregnes som:

$$\hat{y} = \beta_0 + \beta_1 x_0 = 36.4 + 0.059(650) = 74.75$$

Der kan beregnes et konfidensinterval for denne prædiktion. Dette kan gøres som følger, idet alle værdierne tidligere er fundet:

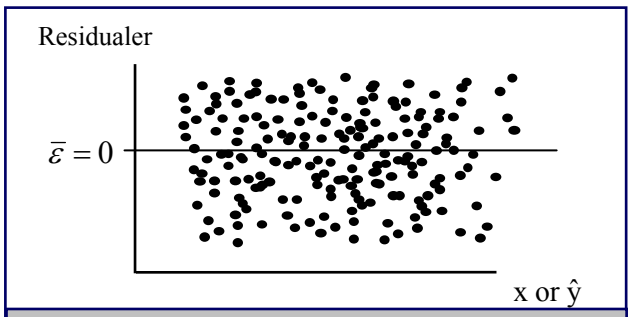
$$\left[\hat{y} \pm t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_x}}\right] \Rightarrow \left[74.75 \pm 2.571(5.96)\sqrt{\frac{1}{7} + \frac{(650 - 400)^2}{280{,}000}}\right] \Rightarrow \left[74.75 \pm 2.571(5.96)\sqrt{0.37}\right]$$

$$\Rightarrow [74.75 \pm 9.27] \Rightarrow [65.48\,;\,84.02]$$

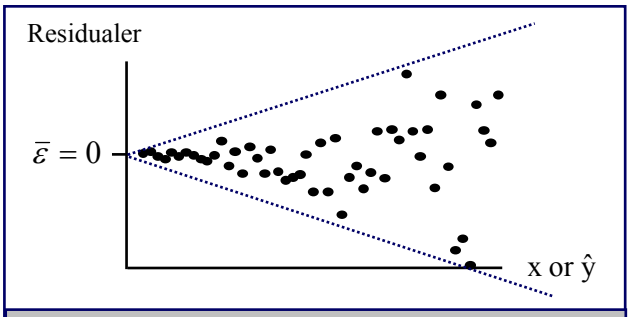
Det vil sige, at der med 95 procents sandsynlighed gælder, at prædiktionen vil falde inden for dette konfidensinterval.

Residualanalyse

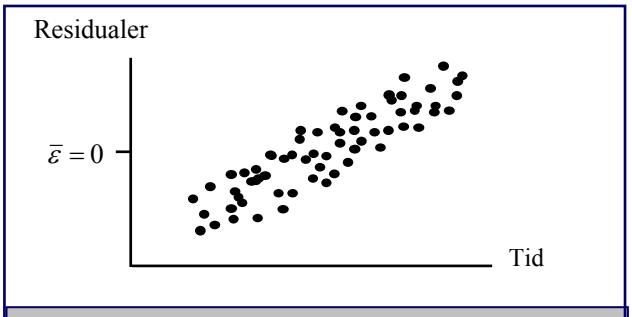
Ved residualanalysen undersøges det, om residualerne overholder de forudsætninger, der ligger til grund for regressionsmodellen. Det vil sige, at de skal være normalfordelte med middelværdi lig nul og konstant varians. Nedenstående viser fire eksempler på diagrammer over residualer.



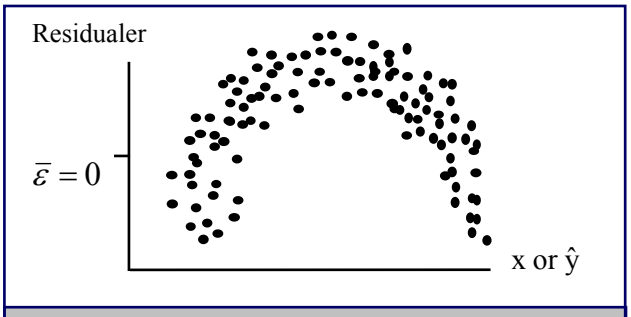
Homoscedasticitet: Residualerne er symmetrisk fordelt omkring nul med fast varians.



Heteroscedasticitet: Variansen er ikke konstant, mens øges med stigende *x*.



Her stiger middelværdien over tid. Modellen er ***ikke-stationær***



Den kurvede form af residualerne indikerer at modellen ikke er lineær.

Situationen øverst til venstre er det korrekte, idet der udvises *hvid støj* (udtrykket stammer fra radiofoniens ungdom), mens de øvrige tre diagrammer viser forskellige former for fejl. I disse tilfælde vil regression give *prædiktionsfejl.* Modellen må da ikke anvendes i nærværende form.

ANOVA i regression

Testet for, om hældningskoefficienten er lig med nul, kan generaliseres til at omfatte alle parametrene. For *k* parametre haves:

H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$
H_1: Mindst én $\beta_1, \beta_2, \ldots, \beta_k$ er forskellig fra nul

Hvis H_0 accepteres siges det, at modellen ikke har nogen mening.

Dette er som et F-test, som udviklet i noterne om ensidet variansanalyse. Der kan opstilles en ANOVA-tabel og udføres et F-test. Tabellen ser uden som følger:

ANOVA-tabel:

Variation	*Kvadrat sum (SS)*	*Frihedsgrader (fg)*	*Middelkvadratsum (MS)*	*F-værdi*
Regression	$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	k	MSR = SSR/k	$F = \frac{MSR}{MSE}$
Residual	$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	n − (k+1)	MSE = SSE/ n−(k+1)	
Total	$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$	n − 1		

F-testet har frihedsgrader lig med $fg_1 = k$ og $fg_2 = n-(k+1)$. I vort eksempel kendes alle værdier med undtagelse af *SSR,* som let kan findes idet såvel *SSE* som *SST* er blevet beregnet. Ved indsættelse fås følgende tabel:

ANOVA-tabel:

Variation	*Kvadrat sum (SS)*	*Frihedsgrader (fg)*	*Middelkvadratsum (MS)*	*F-værdi*
Regression	$SSR = 972.32$	1	MSR = 972.32	$F = \frac{972.32}{35.536} = 27.36$
Residual	$SSE = 177.68$	5	MSE = 35.536	
Total	$SST = 1{,}150$	6		

Testeren har frihedsgrader lig med *(1,5)*. Antages $\alpha = 0.05$ så kan den kritiske værdi findes ved anvendelse af **Statistics Tables.** Her findes at $F_{crit} = 6.61$. Da $27.36 > 6.61$ accepteres H_0 som forventet.

2. Regression på lommeregneren på i Excel

Lad os se ovenpå denne omgang regnerier, hvordan lommeregneren eller Excel kan lette arbejdet!

På **lommeregneren** TI-84/89 indtaster man eksempelvis *x* i registret *L1*, mens *y* kan være i registret *L2*. Dernæst kan man gå videre på 2 måder:

I. Tast STAT → CALC → 4: Linreg(ax+b) → ENTER. Nu er formatet Linreg(ax+b) *L1,L2* → ENTER

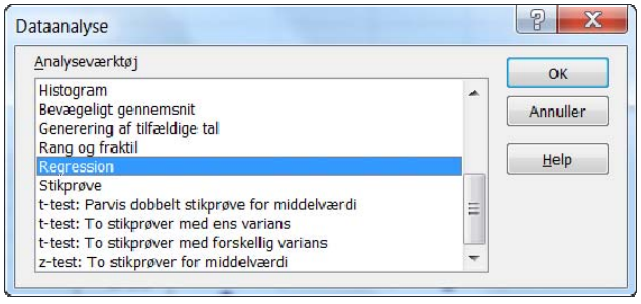
Der fremkommer nu en udskrift indeholdende værdien af parameterne (kaldet a og b), R^2 (determinationskoefficienten) og r (korrelationen).

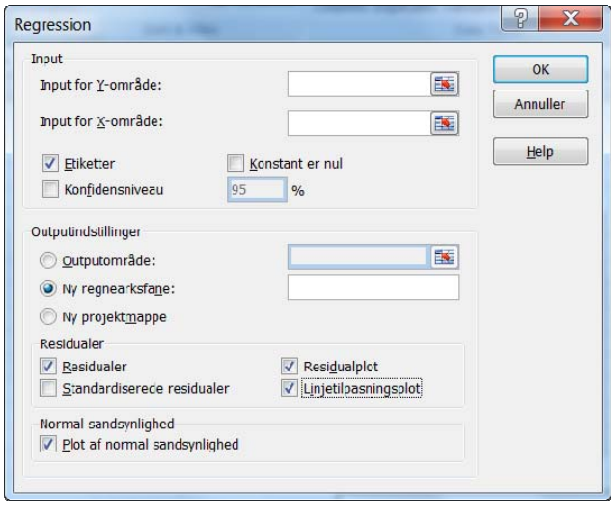
II. Tast STAT → TESTS → F: LinregTTest → ENTER. Nu fremkommer et skærmbillede, hvor man angiver Xlist: *L1* og Ylist: *L2*. Gå nu til CALCULATE → ENTER

Der fremkommer nu en udskrift med frihedsgrader, t-testerens værdi, p-værdien af t-testet, værdien af koefficienterne, standardfejlen, R^2 (determinationskoefficienten) og r (korrelationen).

Option II giver de fleste informationer, og anbefales således. Bemærk det er **vigtigt** at huske i hvilket register, man taster henholdsvis *x* og *y* variablerne. En begrænsning med lommeregneren er, at man alene kan lave en simpel regression med en enkelt *x*-variabel.

Ved anvendelse af **Excel** kan alle beregninger fortages meget enkelt. Åbn programmet og test ***Data/dataanalyse/regression.*** Det skulle give menuen til venstre. Her vælges *regression*, hvorefter menuen til højre fremkommer.





I boksen markeres, som det er angivet. Det vil sige ”etiketter” hvis der er overskrifter med, og der kan markeres for residualer, hvis residual- og linjetilpasningsplot ønskes med taget. Der markeres ”OK”, og følgende udskrift vil fremkomme for vort eksempel:

Regressionsstatistik	
Multiple R	0.92
R-kvadrat	0.85
Justeret R^2	0.81
Standardfejl	5.96
Observationer	7

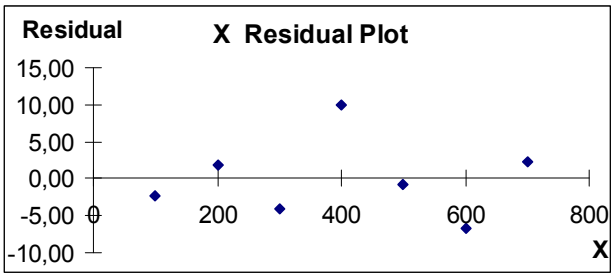
ANOVA

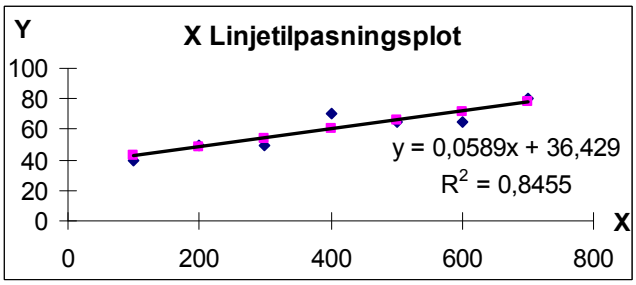
	fg	*SS*	*MK*	*F*	*P-værdi*
Regression	1	972.2	972.32	27.36	0,00
Residual	5	177.68	35.54		
Total	6	1,150.00			

	Koefficient	*Stand.fejl*	*t Stat*	*P-værdi*	*Nedre 95%*	*Øvre 95%*
Skæring	36.43	5.04	7.23	0.00	23.48	49.38
X	0.06	0.01	5.23	0.00	0.03	0.09

Her har jeg selv tilføjet den gule signatur. Endvidere er alle data rundet af til to decimaler. Det letter overblikket gevaldigt, og gør det især let at læse *p-værdierne.* Bemærk, at korrelationen i Excel benævnes ”Multipel R”. I Excel er der tillige en ANOVA-tabel, den justerede R^2-værdi og konfidensintervaller for parametrene i forhold til på lommeregneren.

Residual- og lingetilpasningsplottet ser ud som følger





Ligningen for regressionslinjen har jeg selv tilføjet regressionslinjen. Man højreklikker på et datapar og markerer ”tilføj trendlinje”. Her kan man angive om man ønsker at få medtaget formlen på linjen og R^2.

I Excel er der også mulighed for at udskrive et ”plot af normal sandsynlighed”. Dette diagram anvendes til at undersøge, om materialet opfylder antagelsen om at være normalfordelt. Dette diagram skal helst udvise en ret positiv linje. Yderligere fortolkning af dette diagram ligger udenfor pensum af Statistik II.

3. Multipel regression og læsning af regressionsudskrift

I den multiple regressionsanalyse medtages flere forklarende variable *x*. Regressionsmodellens antagelser er de samme, som er gældende for den simple regressionsmodel.

Modellen kan med *k regressorer* skrives som:

$$y_i = E(y_i) + \varepsilon_i = \beta_0 + \beta_1 x_{1.i} + \beta_2 x_{2.i} + ... + \beta_k x_{k.i} + \varepsilon_i \qquad \text{hvor } i = 1,2, ...,n$$

For at kunne finde parametrene til denne model løses et system med *(k+1)* normalligninger. Det ligger udenfor rammerne af dette kursus at finde disse parametre samt de tilhørende t-værdier, konfidensintervaller mm.

En multipel regressionsmodel kan let estimeres ved anvendelse af **Excel**. I stedet for en enkelt *x*-variabel markeres *alle x*-variable.

I fagbeskrivelsen til Statistik II anføres det, at studenterne skal være i stand til at læse en udskrift fra en multipel regression. Sådanne udskrifter er ofte systematiseret på forskellige måde, så det kan godt være vanskeligt!

Følgende punkter er relevante:

- Er fortegnene på de estimerede parametre som forventet fra den bagved liggende teori eller forventning
- Kan størrelsen tillægges en meningsfuld betydning
- Er parametrene signifikante? Dette undersøges ud fra *p-værdien*. Her gælder følgende:

Svag signifikans:	$p < 0.10$	(90 % niveau α=0.10. Markeres ofte *)
Signifikans:	$p < 0.05$	(95 % niveau α=0.05. Markeres ofte **)
Stærk signifikans:	$p < 0.01$	(99 % niveau α=0.01. Markeres ofte ***)

- Determinationskoefficieterne (R^2 og R^2-justeret) skal være så høj som muligt
- Korrelationen: Så tæt på 1 eller –1 som muligt
- Standardfejlen skal være så lav som muligt, når forskellige modeller sammenlignes

Læsning af en udskrift kan kun læres via øvelse! Lad os se på et eksempel fra en tidligere eksamensopgave.

Eksempel på læsning af en udskrift: Eksamen juni 2011, opgave 5 (20% 4P / 40% 8P)

I den følgende tabel ses resultaterne af en undersøgelse, hvor "skatteundragelse" måles (regressionsmodel 1 og 2) og "kørsel uden billet" (regressionsmodel 3 og 4), som er

forklaret gennem en række variable. Både "skatteundragelse" og "kørsel uden billet" er målt sådan, at et højere tal betyder en større villighed til at begå lovovertrædelserne. Ligeledes ved de andre variable: Højere alder, højere uddannelse, højere indstilling til at overholde lovene etc.

I tabellen nedenfor er vist koefficienterne i forskellige regressionsmodeller. Der er to tal i hver celle.

- Tallet uden parentes er hældningskoefficienterne β til de standardiserede variable
- Tallet i parentes er hældningskoefficienterne β til de ikke standardiserede variable

Prädiktoren	Steuer-betrug (Modell 1)	Steuer-betrug (Modell 2)	Schwarz-fahren (Modell 3)	Schwarz-fahren (Modell 4)
Einstellung zum Delikt	.11** (.10)	.12** (.11)	.11*** (.17)	.09** (.15)
Handlungsintention Ja = 1; Nein = 0	.50*** (.86)	.49*** (.85)	.43*** (1.18)	.39*** (1.06)
Entdeckungswahrschein-lichkeit	-.07* (-.04)	-.08* (-.05)	-.04 (-.04)	-.02 (-.02)
allgemeine Gesetzestreue	.05 (.05)	.01 (.01)	-.07*** (-.12)	-.06 (-.09)
Postmaterialismus (Inglehart-Index)		-.05 (-.04)		-.04 (-.05)
Links-Rechts-Selbst-einstufung		-.01 (-.005)		-.02 (-.01)
Religiosität		-.04 (-.01)		-.06* (-.02)
soziale Schicht		.06 (.08)		.02 (.03)
Geschlecht M = 1; F = 2		-.06 (-.10)		-.11*** (-.26)
Alter		.10** (.007)		-.12*** (-.008)
Bildung		-.06 (-.04)		.08* (.08)
Einkommen		.05 (.00002)		-.01 (-.00001)
korrigiertes R^2	.30 (.33)	.32 (.35)	.25 (.26)	.30 (.30)
N	702	564	1254	887

* = p < .05 ** = p < .01 *** = p < .001

Oversættelse af variablerne er som følger:

Steuerbetrug = Skatteunddragelse	Prædiktoren = Uafhængig variable (β)
Schwarzfahren = Kørsel uden billet	Religiosität = Religiøsitet
Einstellung zum Delikt = Villighed til at lovovertræde	Soziale Schicht = Socialgruppe
Entdeckungwahrscheinlichkeit = Afsløringssandsynligheden	Geschlecht = Køn
Allgemeine Gesetztreue = Indstilling til at overholde love	Alter = Alder
Postmaterialismus = Postmaterialisme	Bildung = Uddannelsesniveau
Links – Rechts – Selbst-Einstufung = Selvplacering v/h skala	Einkommen = indkomst

Svar på følgende spørgsmål ved hjælp af resultaterne i tabellen. Svarene skal begrundes med de anførte tal fra tabellen!

Del 1: (alle svar tillægges vægten 1P eller 5 %)

1. Hvad er den generelle forskel mellem model 1 og 2 samt forskellen mellem model 3 og 4
2. Hvilke variable har signifikant indflydelse på ”skatteunddragelse” og ”kørsel uden kørekort”?
3. Hvor godt forklarer model 2 ”skatteunddragelse” og model 4 ”kørsel uden kørekort”?
4. Hvem har den største intention til ”skatteunddragelse” og ”kørsel uden kørekort” – mænd eller kvinder?

Her kunne man vælge, at gå videre med følgende fire spørgsmål eller subsidiært svar på en anden opgave.

5. Hvilken indflydelse har den stigende ”alder” på ”skatteunddragelse” og ”kørsel uden billet”?
6. Hvilke af de to variable ”alder” og ”afsløringssandsynlighed” har størst effekt på intentionen til ”skatteunddragelse”?
7. Hvad betyder forskellen mellem koefficienterne for ”generel indstilling til at overholde love” mellem model 3 (0.11) og model 4 (0.09)? Hvordan kan man forklare denne forskel?
8. Har religiøsitet indflydelse på begge variable (”skatteunddragelse” og ”kørsel uden kørekort”)?

(Jeg er lidt kritisk overfor oversættelsen. ”Schwarzfahren” kan vel enten betyde, at køre uden kørekort, men vel også at køre uden billet til kollektiv trafik)?

Kommentarer og besvarelse

Indledningsvis bemærker forfatteren til disse noter, at det ikke er normalt at angive parameterestimater udarbejdet efter to forskellige algoritmer (standardiserede og ikke standardiserede regnemetoder). Beregninger af denne karakter hører hjemme i et kursus i økonometri (en blanding af nationaløkonomi og statistik), som ligger fjernt fra rammerne for dette kursus.

Får man en opgave af denne karakter, så tager man det roligt, og fokuserer alene på opgaven! Det vil sige: Læs tabellen grundigt. I en tabel læses tallene *til sidst*!

Modeltypen, som der rapporteres resultater fra i tabellen, er en *logistisk model*. I en sådan regression kan *y*-variablen kun antage to værdier, nemlig nul og ét. Ser man eksempelvis på model 1 om "skatteunddragelse", så kan man have lavet dette eller ej. Det vil sige, at hvis man har unddraget skat, så er værdien af *y* lig med ét, mens den er nul, hvis man ikke har unddraget skat. Det siger sig selv, at hele grundlaget for beregningen er lidt tvivlsomt, da man her erkender en lovovertrædelse.

1)
Hvad er den generelle forskel mellem model 1 og 2 samt forskellen mellem model 3 og 4?

Modellerne 1 og 3 indeholder ikke så mange baggrundsfaktorer, som modellerne 2 og 4. Modellerne 2 og 4 har en højere forklaringsgrad R^2, men er baseret på færre observationer.

2)
Hvilke variable har signifikant indflydelse på "skatteunddragelse" og "kørsel uden kørekort"?

Signifikante variable er alle, som er markeret med en eller flere stjerner.

Her må det være underforstået at man ser på modellerne med alle baggrundsvariabler medtaget. Det vil sige modellerne 2 og 4.

Model 2: "Skatteunddragelse"
Her er følgende variable signifikante:

- Villighed til at lovovertræde (ikke overraskende er denne positiv)
- Handlingsintension – har man gjort dette bevist (positiv effekt)
- Afsløringssandsynligheden (negativ – så man regner med at slippe for straf)
- Alder (positiv – desto ældre desto større sandsynlighed for overtrædelse)

Model 4: "Kørsel uden billet"
Her er følgende variable signifikante:

- Villighed til at lovovertræde (ikke overraskende er denne positiv)
- Handlingsintension – har man gjort dette bevist (positiv effekt)
- Indstilling til at overholde loven (kun i model 3 – går ud i model 4)
- Religion (negativ − det betyder vel, at hvis man er religiøs, så kører man ikke uden billet)?
- Køn (negativ- forfatteren kan ikke tolke dette)!
- Alder (negativ – desto ældre desto mindre sandsynlighed for overtrædelse)

- Uddannelse (positiv – desto mere uddannelse, desto større sandsynlighed for at køre uden billet)

Generelt er model 4 bedre i relation til baggrundsvariablene end model 2.

3)
Hvor godt forklarer model 2 "skatteunddragelse" og model 4 "kørsel uden kørekort"?

Her anvendes R^2. Model 2: $R^2 = .32$ Model 4: $R^2 = .30$

Det vil sige, at model 4 har den bedste forklaringsgrad

4)
Hvem har den største intention til "skatteunddragelse" og "kørsel uden kørekort" – mænd eller kvinder?

Det kan man ikke sige noget om for "skatteunddragelse", da variablen ikke er signifikant! Med hensyn til "kørsel uden kørekort", så er koefficienten negativ signifikant. Da mand = 1 og kvinde = 2, så må det være mænd, der har den største tilbøjelighed til overtrædelsen, da værdien af dummyvariablen for mænd er mindre end for kvinder.

5)
Hvilken indflydelse har den stigende "alder" på "skatteunddragelse" og "kørsel uden billet"?

I begge modeller er parameteren signifikant, men med modsat fortegn. Under "skatteunddragelse" stiger sandsynligheden med alderen, mens det omvendte er tilfældet med "kørsel uden kørekort" (sidstnævnte kunne skyldes, at yngre ofte end ældre kører i kollektiv transport uden gyldig billet).

6)
Hvilke af de to variable "alder" og "afsløringssandsynlighed" har størst effekt på intentionen til "skatteunddragelse"?

Model 2 betragtes, og begge parametre er signifikante. Alder må have den største effekt, da den numeriske værdi af parameteren her er størst (0.10)

7)
Hvad betyder forskellen mellem koefficienterne for ”generel indstilling til at overholde love” mellem model 3 (0.11) og model 4 (0.09)? Hvordan kan man forklare denne forskel?

Koefficienten er i begge modeller signifikant. Effekten (tilbøjeligheden til at”køre uden kørekort”) i model 3 er større end i model 4. Model 3 indeholder ikke de samme baggrundsparametre, som model 4.

8)
Har religiøsitet indflydelse på begge variable (”skatteunddragelse” og ”kørsel uden kørekort”)?

Religiøsitet har alene indflydelse på ”kørsel uden kørekort” (det er der vist blevet svaret på tidligere), da variablen her er signifikant.