**Excel-øvelser i**

**statistikk**

**av**

**Peer Andersen**

**© Peer Andersen 2010**

# Statistikk

Vi skal i denne bolken se på hvordan Excel kan brukes i forhold til enkle statistiske betraktninger. Excel er velegnet til statistisk arbeid, og det finnes en rekke funksjoner som kan anvendes også i forhold til mer avanserte betraktninger. Vi skal i denne omgang nøye oss med noen forholdsvis enkle øvelser. Vi skal først se på en øvelse der vi kartlegger antall søsken som elevene i en klasse har. Dette er et eksempel på det vi kaller en diskret fordeling. Det vil si at vi bare kan få et begrenset antall utfall. I teorien kan en person ha svært mange søsken, men i en normal situasjon er det et begrenset antall, og vanligvis under 10. Vi skal sortere dataene våre og fremstille dem på en oversiktlig måte ved å bruke stolpediagram.

I den neste øvelsen skal vi se på hvor mange minutter den enkelte elev bruker på veien til skolen. Dette er et eksempel på en kontinuerlig fordeling. Det vil si at det er uendelig mange utfall, ingen elever bruker eksakt like lang tid til skolen. Vi har for enkelthets skyld rundet av til hele minutter. Når vi skal fremstille slike data er det vanlig å klassedele materialet, det vil si vi teller opp hvor mange som f. eks. bruker fra 1 til 5 minutter til skolen, hvor mange som bruker fra 6 til 10 minutter osv. Når vi skal fremstille dette, er det vanlig å bruke en variant av stolpediagrammet som vi kaller histogram. Den tredje øvelsen vi skal se på handler om hvordan statistikk kan misbrukes og hvordan vi kan manipulere en fremstilling slik at vi får frem akkurat det resultatet som vi ønsker oss, men som ikke nødvendigvis gir et korrekt bilde av virkeligheten. Den fjerde øvelsen i dette kapittelet handler om ulike sentralmål og spredningsmål. Vi skal se hvordan Excel kan brukes til å beregne størrelser som gjennomsnitt, median og standardavvik. Til slutt skal vi se på en øvelse om regresjon. Regresjon går kort fortalt ut på at vi ut i fra noen observasjoner kan finne en funksjon som er tilpasset disse dataene. Dette er meget nyttig i forhold til å lage prognoser for fremtidig utvikling.

Statistikkøvelsene er ikke så veldig vanskelige teknisk sett. Vi bruker helt grunnleggende funksjoner i Excel. Vanskelighetsgraden er slik at øvelsene burde la seg gjennomføre i en skoleklasse.

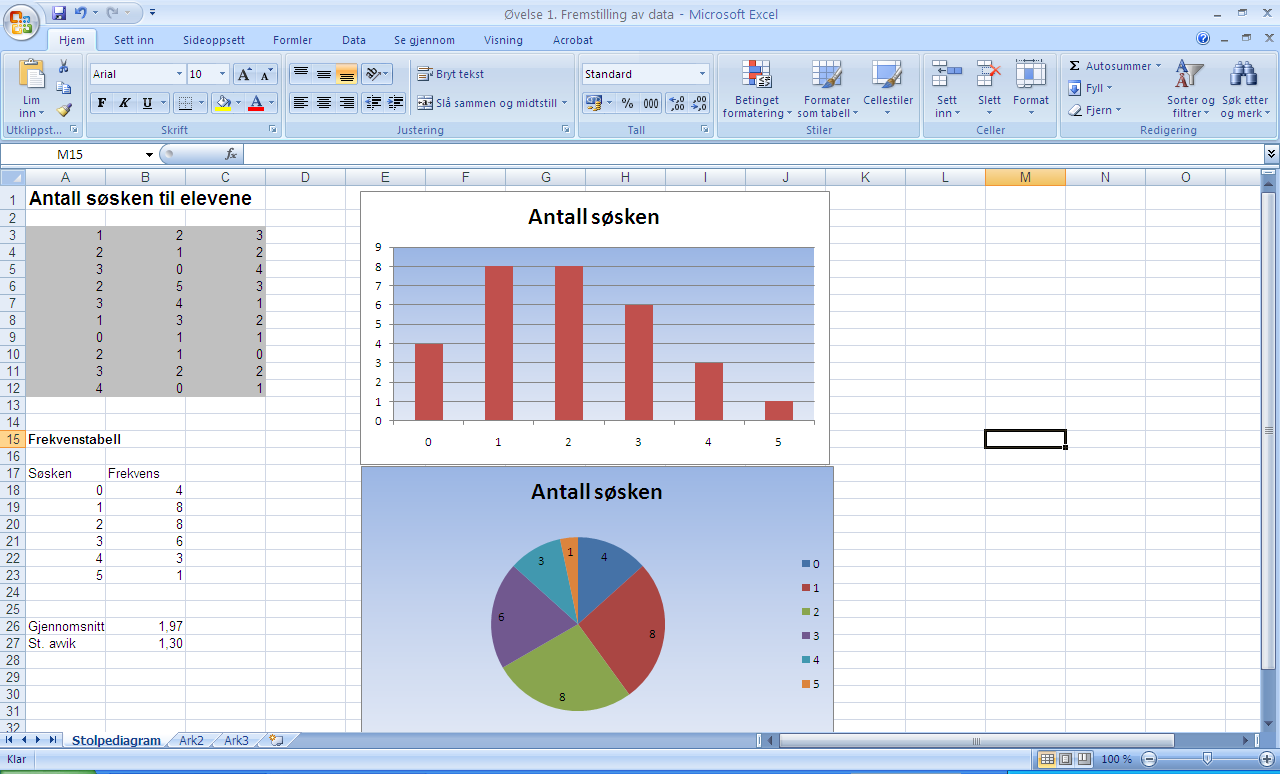
## Øvelse 1: Fremstilling av data

Vi skal i denne øvelsen se på hvordan vi kan systematisere data i en frekvenstabell. Vi skal deretter se på hvordan dataene kan fremstilles i stolpediagram og sektordiagram.

Vi skal ta utgangspunkt i en klasse med 30 elever og kartlegge hvor mange søsken hver enkelt elev har. Vi bruker dataene som er gitt i tabellen under, men du kan gjerne gjøre en undersøkelse i klassen og bruke dataene derfra i stedet for.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Antall søsken til elever i en klasse** | | | | | | | | | |
| 1 | 2 | 3 | 2 | 3 | 1 | 0 | 2 | 3 | 4 |
| 2 | 1 | 0 | 5 | 4 | 3 | 1 | 1 | 2 | 0 |
| 3 | 2 | 4 | 3 | 1 | 2 | 1 | 0 | 2 | 1 |

Vi skal legge inn dataene i et Excel regneark og deretter behandle dem der. Når regnearket vårt er ferdig, kan det se ut omtrent som vist under.



Konstruksjon av regnearket

Vi skal nå se på hvordan vi kan lage dette regnearket. Det er fornuftig å starte med å skrive en liten tittel i rute A1, f. eks Antall søsken til elevene. I feltene som er gråfarget legger du så inn antall søsken til den enkelte elev. Det neste vi skal gjøre er å lage en frekvenstabell der vi kartlegger hvor mange som har ingen søsken, ett søsken, to søsken osv. I eksempelet over er frekvenstabellen plassert i rute A17 til B23. I kolonne A skriver vi inn antall søsken, mens i kolonne B skal maskinen beregne hvor stor frekvensen er. For å beregne frekvensen må vi benytte oss av en funksjon som heter ANTALL.HVIS. Dette er en funksjon som søker gjennom en gitt datamengde og teller opp antall elementer som oppfyller et gitt vilkår. Vi ser først på hvordan vi kan beregne hvor mange som har 0 søsken. Det skal vi gjøre i celle B18. Flytt musen til B18 og klikk deretter på funksjonsveiviseren. Der leter du deg frem til funksjonen Antall.Hvis. Når du åpner denne får du frem følgende vindu



Du kan nå flytte musen til det hvite feltet etter Område. Merk deretter hele det grå feltet. Når det er gjort flytter du musen til det hvite feltet etter Vilkår. Skriv der inn tallet 0. Vinduet skal nå se ut som dette



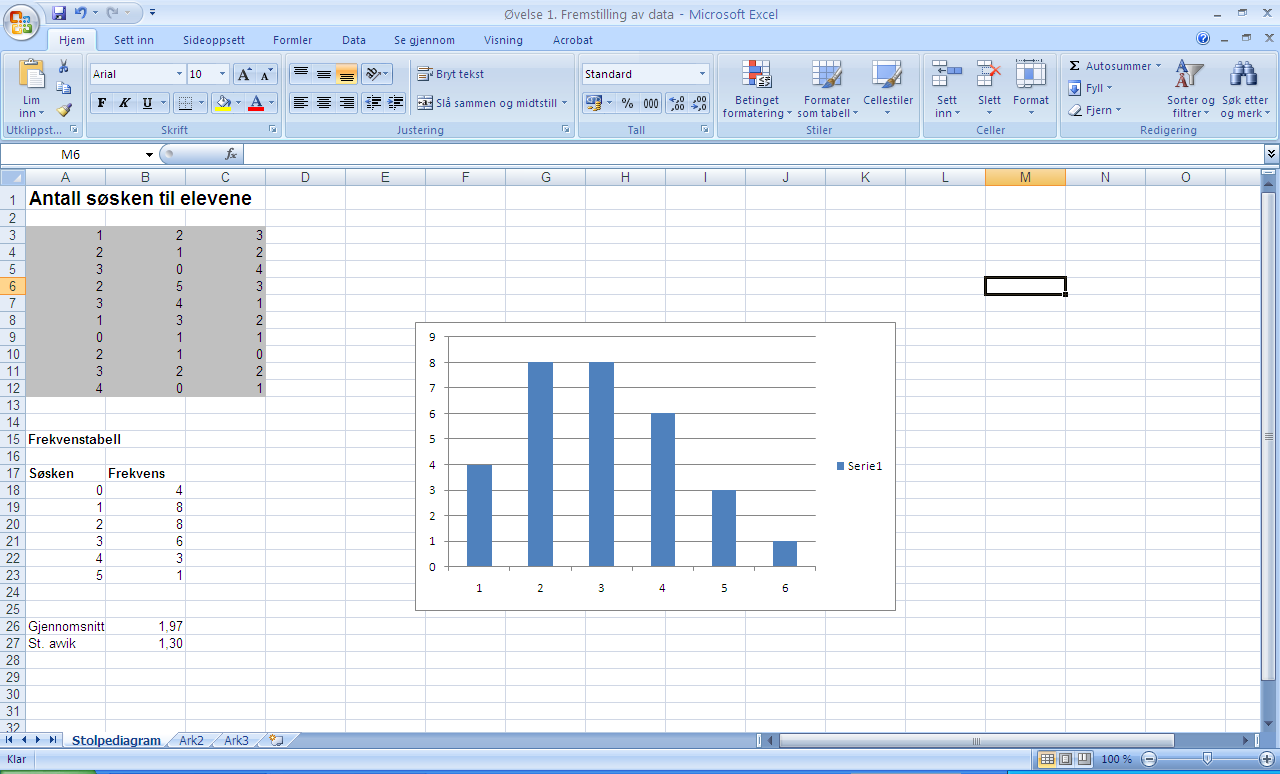
Det Excel da gjør er at den søker gjennom feltene fra A3 til C12, altså det som er merket med grått. Deretter teller den opp antall elementer som er lik 0. På tilsvarende måte kan du bruke funksjonen til å telle opp hvor mange som har ett søsken, to søsken osv.

Noen vil kanskje synes det er tungvint å gjenta denne prosedyren 6 ganger. Det finnes en måte å unngå det på ved å bruke kopiering. Vi må da modifisere litt på det vi gjorde i rute B18. Du kan flytte musen til rute B18 og klikke på knappen. Du får da opp vinduet fra i sted. (Det som er vist over). Vi endrer nå det som står i feltet etter Område til

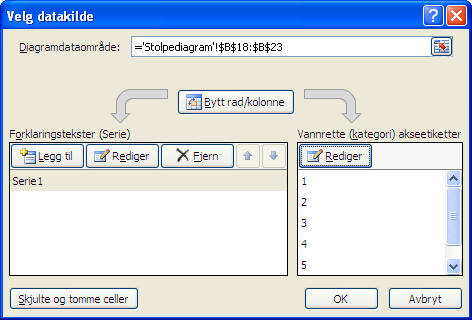
$A$3:$C$12

På vilkår kan du skrive inn A18 istedenfor 0. Denne formelen kan kopieres nedover til de cellene B19 til B23. Forskjellen på denne formelen i forhold til den opprinnelige er at vi har satt på noen dollartegn. Dersom vi plasserer dollartegn før og etter bokstavene i cellereferansene, kan formelen kopieres uten at cellereferansene endres. Dersom vi utelater dollartegnet, vil cellereferansene endres. Du legger også merke til at vi istedenfor 0 har skrevet inn A18. Tallet som er i rute A18 er lik 0 så resultatet blir det samme. Det som imidlertid skjer når vi kopierer formelen til rute B19, er at A18 blir til A19, og den henter da ut tallet 1 istedenfor 0.

Når vi har konstruert frekvenstabellen, er det i grunnen ganske enkelt å fremstille dataene i stolpediagram. Du starter med å merke cellene B18 til B23. Gå til Sett inn på menyen velg stolpe. Velg diagrammet øverst til venstre. Diagrammet du har fått frem ser antagelig ut omtrent som det som er vist på neste side.



Vi legger merke til at enhetene på stolpene ikke stemmer. Det skal vi først rette på. Måten vi gjør det på er at du først høyreklikker på en av stolpene. Velg der Merk data. Klikk på Rediger til høyre i vinduet du får opp.



Merk deretter A18 til A23 i det vinduet du deretter får opp. Klikk så på ok. Klikk også på ok i det neste vinduet. Nå skal enhetene være riktige. Du kan også fjerne Serie1 ved å klikke på den og trykke delete. Du kan også endre farger på området og på stolpene. Hvis du vil endre farge på stolpene høyreklikker du på en av stolpene og velger deretter Formater dataserie. Velg der Fyll og deretter f. eks heldekkende. Til slutt finner du en passende farge. Hvis du vil endre farge på selve diagramområdet klikker du en eller annen plass i diagrammet utenom stolpene. Velg deretter Formater Tegnområde. Velg igjen Fyll og velg et passende alternativ. Jeg har i mitt eksempel valgt gradert fyll.

Vi skal også lage et sektordiagram. Det gjøres etter samme mal som stolpediagrammet. Merk først rute B18 til B23. Gå til Sett inn og velg sektordiagrammet. Også her ser du at enhetene ikke stemmer. Følg samme prosedyre som i sted ved å gå til Merk data for å endre det. Regnearket bør nå se ut som det som er vist i innledningen til denne oppgaven. Hvis du vil ha inn verdier på de forskjellige kakestykkene kan du høyreklikke en eller annen plass i diagrammet og deretter velge Legg til dataetiketter. Du kan tilpasse disse ved etterpå å velge Formater dataetiketter. Her kan du velge f. eks å oppgi verdiene i prosent etc.

Excel kan enkelt beregne gjennomsnitt og standardavvik for oss. Det finnes også en rekke andre funksjoner i Excel, men de skal vi ikke benytte nå. I rute A26 og A27 kan du skrive inn henholdsvis Gjennomsnitt og St. avvik. I rute B26 skal vi bruke en funksjon som heter

=GJENNOMSNITT(A3:C12)

Denne funksjonen beregner gjennomsnittet av tallene som ligger i rute A3 til C12. På tilsvarende måte kan standardavviket beregnes. Da kan vi bruke funksjonen

=STDAVP(A3:C12)

Denne beregner standardavviket av tallene som ligger i rute A3 til C12.

I stedet for å skrive inn funksjonene for gjennomsnitt og standardavvik kan du selvsagt bruke funksjonsveiviseren. Resultatet blir det samme.

Oppgave

a) Studer regnearket du nettopp har laget. Hva synes du gir den beste fremstillingen, stolpediagrammet eller sektordiagrammet?

b) En av de store fordelene er at en kan endre på dataene som er lagt inn, og at regnearket automatisk oppdateres. Prøv å endre på noen av rutene i det grå feltet. Hva skjer med frekvenstabellen og diagrammene? Hva skjer med gjennomsnittet og standardavviket?

## Øvelse 2: Klassedelt materiale – histogram

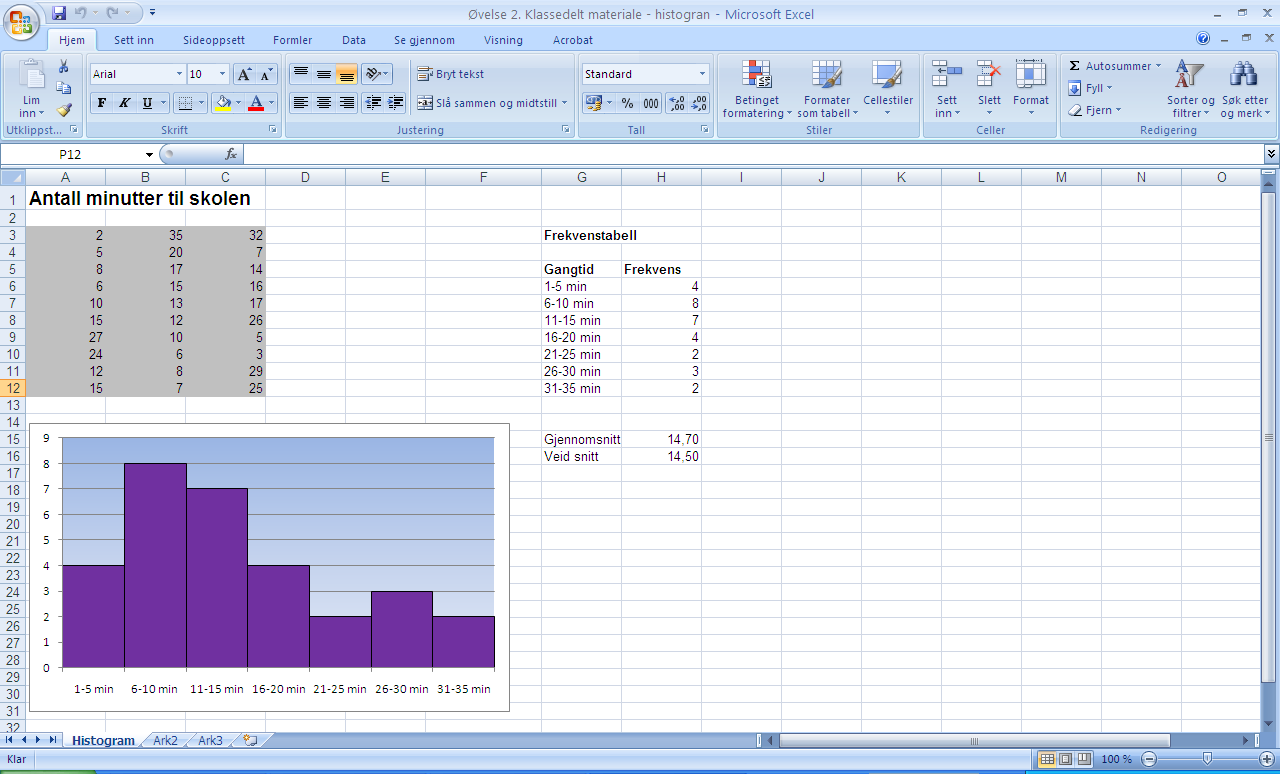
I øvelse 1 så vi på et datamateriale der vi bare hadde noen få forskjellige utfall, fra 0 til 6 søsken. I andre tilfeller er datamaterialet det vi kaller kontinuerlig fordelt, det vil si at vi kan få alle mulige typer utfall, og i prinsippet at alle utfallene er forskjellige. Et eksempel på dette kan være tider i langrennsløp, lengden på blader på trær, vekt på fisk etc. I slike sammenhenger vil det være meningsløst å tegne opp et stolpediagram fordi frekvensen i prinsippet vil være 1 for alle utfall. I et slikt datamateriale vil det være mer fornuftig å klassedele dataene og tegne opp et histogram. Det skal vi se nærmere på i denne øvelsen.

Vi skal også her ta utgangspunkt i en klasse med 30 elever og kartlegge hvor mange minutter den enkelte elev bruker på å gå til skolen. Vi har i denne øvelsen brukt tallene fra tabellen under, men du kan gjerne lage en undersøkelse i egen klasse og bruke disse dataene istedenfor. Vi har for enkelthets skyld rundet av til nærmeste hele minutt.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Antall minutter til skolen** | | | | | | | | | |
| 2 | 5 | 8 | 6 | 10 | 15 | 27 | 12 | 24 | 15 |
| 35 | 20 | 17 | 15 | 13 | 12 | 10 | 6 | 8 | 7 |
| 32 | 7 | 14 | 16 | 17 | 26 | 5 | 3 | 29 | 25 |

Konstruksjon av regnearket

Vi skal nå se på hvordan vi kan lage dette regnearket. Når regnearket vårt er ferdig, kan det se ut omtrent som vist under.



Det er fornuftig å starte med å skrive en liten tittel i rute A1, f. eks Antall minutter til skolen. I feltene som er gråfarget legger du så inn antall minutter som den enkelte bruker til skolen.

Det neste vi skal gjøre er å lage en frekvenstabell der vi grupper dataene. Vi lager oss noen kategorier, f. eks en for 1-5 minutter gangtid, en for 6–10 minutters gangtid, en for 11–15 minutters gangtid osv. Vi skal nå bruke ANTALL.HVIS funksjonen til å få laget frekvenstabellen. Dette blir noe annerledes enn for datamaterialet som ikke er klassedelt. Vi kan starte med å skrive inn de ulike klassene i feltene G6 til G12. I kolonne H skal regnearket beregne frekvensen. Vi starter med kategorien 1–5 minutters gangtid som blir litt ulik de andre. I rute H6 kan vi bruke formelen

=ANTALL.HVIS(A3:C12;”<=5”)

Det Excel da gjør er at den søker gjennom feltene fra A3 til C12, altså det som er merket med grått. Deretter teller den opp antall elementer som er mindre eller lik 5. Vi finner da antallet som har mellom 1 og 5 minutters gangtid. Du kan selvsagt bruke funksjonsveiviseren istedenfor.

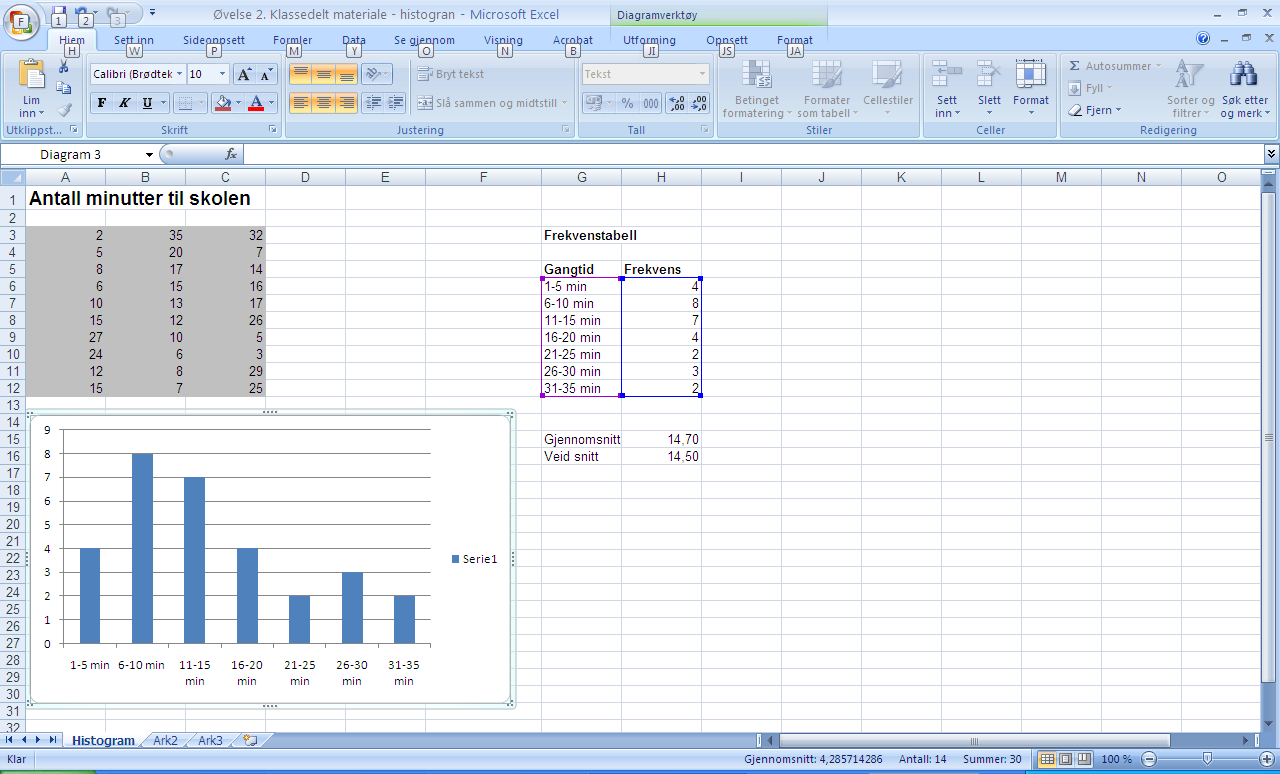
Når vi nå skal se på klassen fra 6–10 minutter, må vi gjøre det på en litt annen måte. Vi bruker først ANTALL.HVIS funksjonen til å finne hvor mange som har 10 minutter eller kortere gange, og så trekker vi fra de som har 5 minutter eller mindre gangtid. Funksjonen

=ANTALL.HVIS(A3:C12;”<=10”) – ANTALL.HVIS(A3:C12;”<=5”)

gir oss da antallet som har mellom 6 og 10 minutters gangtid. Resten av tabellen fylles ut på tilsvarende måte. Det kan lønne seg å bruke dollartegn på området som ANTALL.HVIS funksjonen søker gjennom, så istedenfor A3:C12 i formelen kan vi skrive $A$3:$C$12. Da kan funksjonen kopieres nedover, og alt vi trenger å gjøre er å modifisere ”<=10” og ”<=5” slik at funksjonen samsvarer med det som vi skal finne frekvensen av.

Utfylling av frekvenstabellen kan gjøres mer elegant ved å bruke FREKVENS-formelen istedenfor. Dette er det vi kaller en matriseformel. Matriseformler faller utenfor rammene av denne boken, og det overlates til de spesielt interesserte å studere denne.

Når dette er gjort, skal vi lage histogrammet. Først lager du et stolpediagram etter samme mal som i Øvelse 1. Merk cellene G6 til H12, og klikk på Sett inn og deretter på Stolpe. Du ser nå at vi har fått opp et stolpediagram. Vi legger også merke til at vi her har fått riktige enheter på stolpene. Årsaken til at den betrakter den første kolonnen som verdier på x-aksen, er at vi her har tekst i kolonnen. I det øyeblikket vi har tall i kolonnen betrakter Excel den som noe som skal fremstilles grafisk og den setter da verdiene på x-aksen fra 1 og oppover. Regnearket bør nå se ut omtrent som vist på under.



Vi skal nå fikse litt på diagrammet slik at vi får et histogram. Flytt musen til en av de blå stolpene og klikk med høyre musetast. Velg formater dataserie og velg deretter Seriealternativer. I vinduet du da får opp kan du velge størrelsen på mellomrommet. Sett denne til 0%. Velg deretter Kantlinjefarge. Merk av for heltrukken linje og velg svart som farge. Du skal nå ha fått frem et histogram. Du kan eventuelt selv velge andre farger eller skyggelegging om du ønsker det. Regnearket ditt bør nå se ut omtrent som det som er vist i starten på denne øvelsen.

Vi skal også her beregne gjennomsnittet. Det skal vi gjøre på to måter. Den ene er å beregne det på vanlig måte ved å legge sammen alle 30 observasjonene og dele på 30. Vi finner da det vi kaller aritmetisk middelverdi. Det kan gjøres ved å bruke funksjonen GJENNOMSNITT slik du brukte den i Øvelse 1. Gjør dette, og sett svaret litt under frekvenstabellen. Den andre måten er det vi kaller veid gjennomsnitt. Vi tar da utgangspunkt i klassemidtpunktet og ganger det med frekvensen. Dette gjøres for alle klassene, og til slutt legger vi sammen svarene og deler på 30. Det finnes så vidt jeg vet ingen funksjon som gjør dette i Excel, så vi må lage formelen selv. Formelen

=(3\*H6+8\*H7+13\*H8+18\*H9+23\*H10+28\*H11+33\*H12)/30

skulle imidlertid gi oss det ønskede resultat.

Oppgave

a) Studer regnearket du nettopp har laget. Diskuter kort hvorfor det er mer fornuftig å bruke et histogram enn et stolpediagram i situasjoner som dette.

b) Gå inn i det grå feltet og endre på noen av tallene. Studer hva som skjer med frekvenstabellen og histogrammet.

c) Vi ser at det er små avvik mellom den aritmetiske middelverdien og det veide gjennomsnittet. Hva er årsakene til dette? Vil det alltid være tilfelle for et klassedelt materiale? Prøv gjerne å legge inn nye tall i det grå feltet for å studere dette.

## Øvelse 3: Misbruk av statistikk

Vi skal i denne øvelsen se litt nærmere på hvordan statistikk kan misbrukes og manipuleres slik at vi får frem det vi ønsker. Hensikten med øvelsen er ikke at vi skal trene oss på å misbruke statistikk, men snarere at vi som lærere bør være klar over farene, og hvor enkelt det egentlig er å misbruke statistikk. Et klassisk eksempel på misbruk av statistikk stod en profilert Høyre-politiker for i forbindelse med valgkampen i 1985. Diskusjonen dreide seg om hvor mange sykehusinnleggelser det hadde vært per 1000 innbygger under forskjellige regjeringer. Høyre-politikeren viste under en TV-debatt frem et diagram som kunne gi inntrykk av at det hadde vært dobbelt så mange innleggelser når Høyre satt med makten i forhold til når Arbeiderpartiet satt med makten. Sannheten er at det hadde vært marginalt flere innleggelser i Høyre-perioden i forhold til Arbeiderparti-perioden. Denne saken vakte en del reaksjoner, og problematikken ble kommentert i landets aviser. Vi skal i denne øvelsen se på en lignende problemstilling.

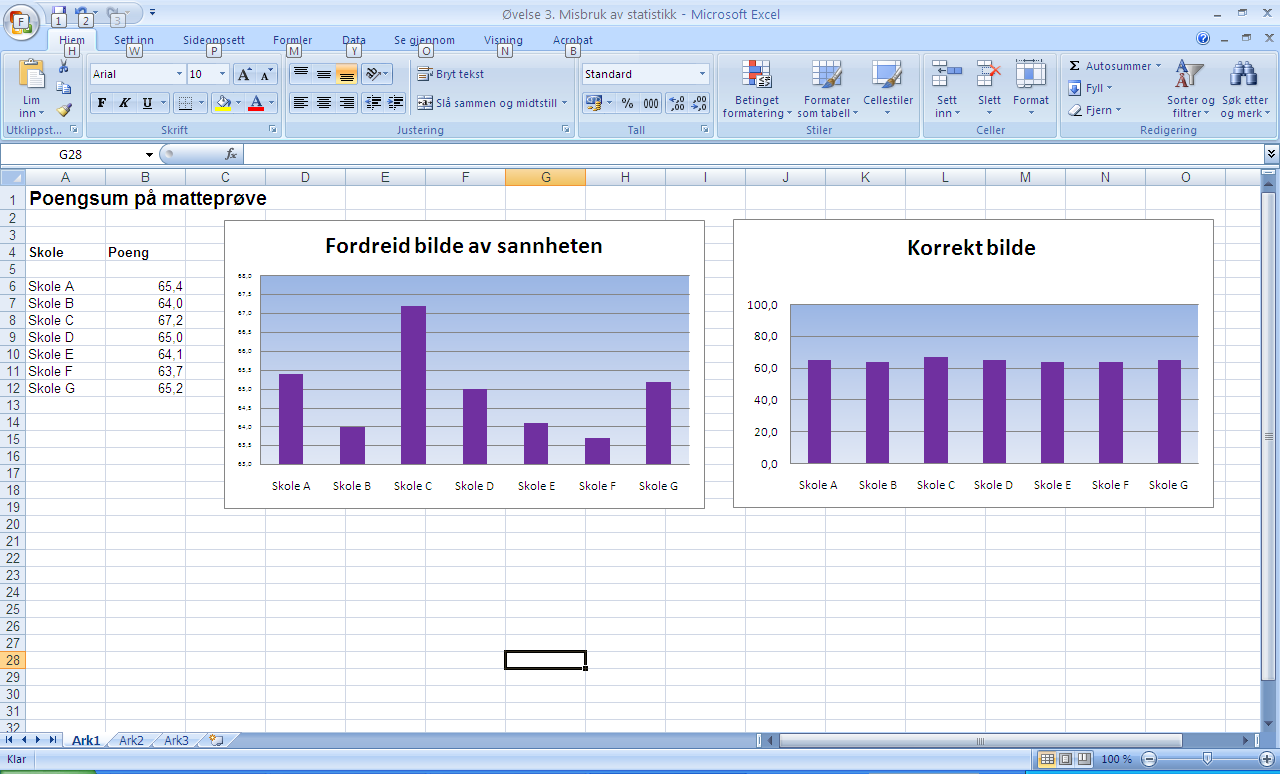
Vi skal studere et eksempel der vi tar utgangspunkt i en matematikkprøve som er gitt til alle 9. klassene i en kommune. Maks poengsum på prøven er 100 poeng, og gjennomsnittscoren for de 7 skolene i kommunen er vist i tabellen under

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Skole** | A | B | C | D | E | F | G |
| **Score** | 65,4 | 64,0 | 67,2 | 65,0 | 64,1 | 63,7 | 65,2 |

Som vi ser så er det skole C som kommer best ut, men forskjellene er heller små. Rektor på skole C er naturlig nok fornøyd med at hennes skole er kommet best ut av det. Men hun er ikke bare fornøyd med å være best, hun vil gjerne fremheve at hennes skole er svært mye bedre enn de andre. Vi skal nå lage et regneark der vi først skal lage et diagram som tilsynelatende viser at skole C er mye bedre enn de andre og etterpå skal vi lage et diagram som gir et mer korrekt bilde av virkeligheten.

Konstruksjon av regnearket

Vi skal konstruere et regneark som ser ut omtrent som det vist under.



Vi starter med å lage en liten tittel, og legger deretter skolenavnene i kolonne A og tilhørende poengsummer i kolonne B.

Vi skal nå lage diagrammet der vi skal fremheve skole C som mye bedre enn de andre. Start med å merke feltene der skolenavnene og poengsummene er gitt, i mitt tilfelle rutene A6 til B12. Klikk på Sett inn og Stolpe. Velg diagrammet øverst til venstre. For å få vekk ruten med serie 1 høyreklikk du på den og klikker på Fjern.

Vi har nå fått frem et diagram som fremhever skole C som betydelig bedre enn de andre. Vi kan imidlertid gå enda lenger. Høyreklikk på et av tallene på y-aksen og velg Formater akse. Velg så Skala, og sett minimum til 62 og maksimum til 68. Ta også vekk hakene foran minimum og maksimum.

Vil vi gå enda lenger kan vi sette minimum til 63. Klikk deretter på ok. Tallene på y-aksen er ganske store og det er lett og se hvilken poengsum de ulike skolene har oppnådd. Disse tallene kan forminskes slik at de blir vanskeligere å lese. Høyreklikk igjen på et av tallene på y-aksen og velg skrift. Sett skriftstørrelsen til 4. Gjør også det samme på x-aksen, men vi kan ha litt større skrift her, f. eks 8 slik at skolenavnet blir leselig. Diagrammet ditt bør nå se ut omtrent som det øverste diagrammet til venstre i eksempelet mitt.

Vi skal nå konstruere det andre diagrammet som gir et mer korrekt bilde av virkeligheten. Vi starter igjen med å merke feltene fra A6 til B12 og klikker på diagramveiviseren. Følg deretter samme fremgangsmåte som ved forrige diagrammet bare at du setter inn en annen tittel, f. eks Korrekt bilde.

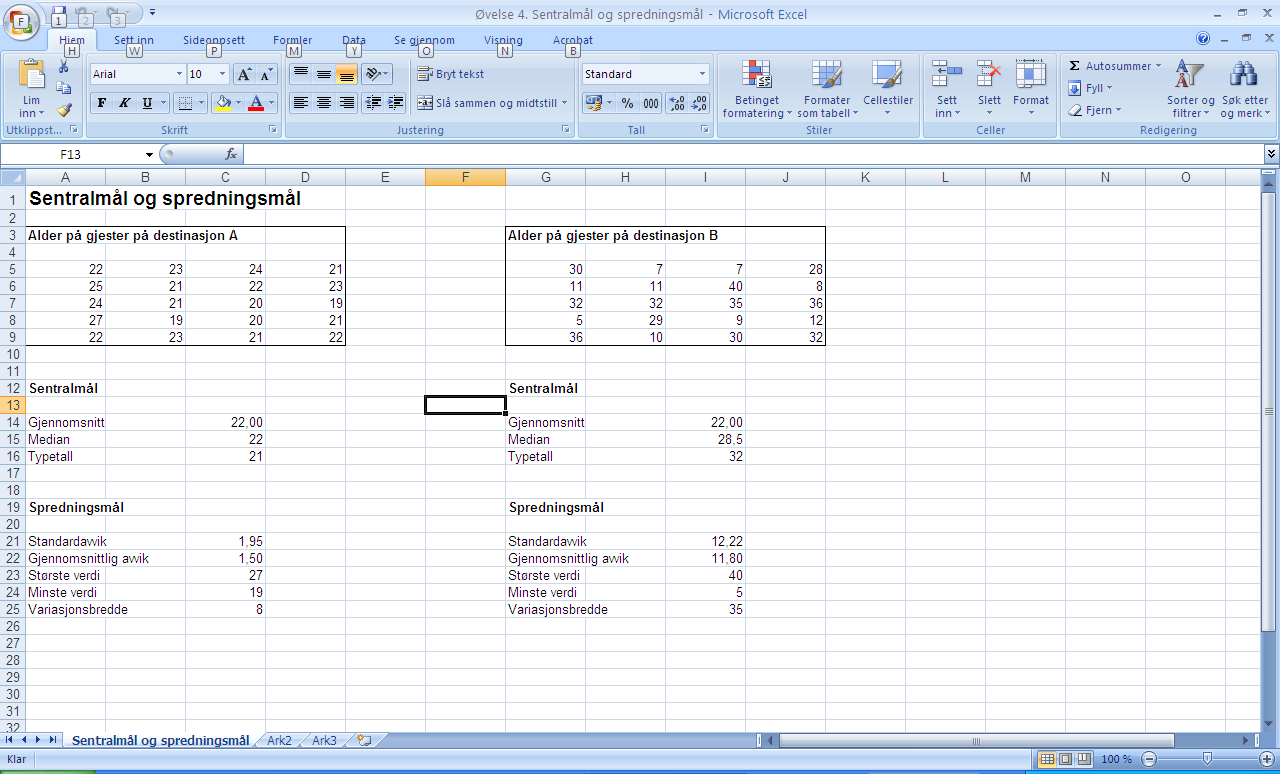
For at dette diagrammet skal gi oss et riktig bilde må vi justere på aksene også her. Høyreklikk på et av tallene på y-aksen og velg formater akse. La nå minimum være 0 og maksimum være 100. Sørg for at det ikke er hake før disse verdiene. Klikk på ok. Fjern også her ruten serie 1. Plasser diagrammene slik at de kommer ved siden av hverandre eller under hverandre.

## Øvelse 4: Sentralmål og spredningsmål

Vi skal i denne øvelsen se på ulike sentralmål og spredningsmål. Typiske sentralmål er gjennomsnitt og median. Standardavvik er et eksempel på spredningsmål. Vi skal se nærmere på hvordan Excel kan brukes til å beregne disse målene, og vi skal også se på nytten av å oppgi både et sentralmål og et spredningsmål i et statistisk materiale. Vi skal i denne øvelsen ta utgangspunkt i to forskjellige syden-destinasjoner. Vi tenker oss at det er 20 gjester fra et bestemt reiseselskap på begge stedene. I tabellen under har vi oppgitt alderen til gjestene på de to destinasjonene.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Destinasjon A** | | | |  | **Destinasjon B** | | | |
| 22 | 23 | 24 | 21 |  | 30 | 7 | 7 | 28 |
| 25 | 21 | 22 | 23 |  | 11 | 11 | 40 | 8 |
| 24 | 21 | 20 | 19 |  | 32 | 32 | 35 | 36 |
| 27 | 19 | 20 | 21 |  | 5 | 29 | 9 | 12 |
| 22 | 23 | 21 | 22 |  | 36 | 10 | 30 | 32 |

Vi skal nå behandle dette i Excel og se på ulike sentral- og spredningsmål. Regnearket vi skal lage, skal se ut omtrent som vist under.

****

Vi skal starte med å legge alderen inn i rutene med ramme rundt. Alderen til gjestene på destinasjon A kan du legge inn i feltet fra A5 til D9. Alderen til gjestene på destinasjon B kan du legge inn i feltet fra G5 til J9. Legg også til passende overskrifter.

Vi har flere typer sentralmål. De meste vanlige er aritmetisk gjennomsnitt, median og typetall. Aritmetisk gjennomsnitt er det vi vanligvis kaller for gjennomsnitt eller middelverdi. Denne finnes ved å legge sammen alderen til alle personene og dele på antall personer. Median finnes ved å ordne tallene i stigende rekkefølge og deretter plukke ut det midterste tallet. Typetallet er den observasjonen som opptrer flest ganger. Med så lite tallmateriale som vi har lar dette seg greit regne ut for hånd. For større datasett er det imidlertid en fordel å bruke datamaskin. Excel regner enkelt ut både gjennomsnitt, median og typetall for oss. Vi skal nå se hvordan dette kan gjøres.

I cellene A12 og G12 kan du skrive inn en liten overskrift, f. eks Sentralmål. I rute A14 og G14 skriver du Gjennomsnitt, i A15 og G15 skriver du Median og i rute A16 og G16 skriver du Typetall. I C- og I-kolonnen skal vi regne ut disse verdiene. Flytt musen til rute C14. Bruk funksjonsveiviseren til å finne frem funksjonen som heter GJENNOMSNITT. Klikk på denne og merk området som du vil ha gjennomsnittet av. I vårt tilfelle er det A5 til D9. Klikk så på ok. Flytt musen til C15 og regn ut Medianen ved å bruke funksjonen som heter MEDIAN. Til slutt skal vi finne typetallet. Flytt først musen til C16. Funksjonen som finner typetallet heter i Excel MODUS. Bruk denne for å finne typetallet. Når dette er gjort, gjør du tilsvarende for gjestene ved destinasjon B. Regnearket bør nå se ut som øverste del av regnearket på foregående side.

Oppgave

a) Hva slags type gjester tror du har reist til destinasjon A, og hvilke gjester tror du har reist til destinasjon B?

b) Hvilke av sentralmålene synes du gir en best beskrivelse av alderen til gjestene på de to destinasjonene? I hvilke tilfeller er det hensiktsmessig å bruke de ulike sentralmålene?

c) Hva blir gjennomsnittlig alder til de to gruppene? Er gjennomsnitt alene et godt nok mål for å si noe om aldersammensetning i en gruppe? Begrunn.

Vi skal nå se på hvordan Excel kan brukes til å beregne standardavvik og andre typer spredningsmål. Standardavvik er forholdsvis krevende å regne ut for hånd, og selv for relativt små datamengder er det en fordel å bruke datamaskin til det. Standardavvik er det mest brukte spredningsmålet og kan beregnes ved å bruke følgende formel.

der er gjennomsnittet for populasjonen og er antallet i populasjonen. I Excel finnes imidlertid denne funksjonen, så vi slipper å skrive inn hele formelen. Gå til rute C21 og åpne funksjonsveiviseren. Finn så en funksjon som heter STDAVP. Klikk på denne og merk datasettet som du vil beregne standardavviket for. I vårt tilfelle er det A5 til D9. Klikk så på ok.

Et annet spredningsmål som også brukes, er gjennomsnittlig absoluttavvik. Gjennomsnittlig absoluttavvik finnes ved å bruke følgende formel:

der er gjennomsnittet for populasjonen og er antallet i populasjonen. Excel har også en funksjon for å beregne gjennomsnittlig avvik. Gå til rute C22 og let deg frem til funksjonen GJENNOMSNITTSAVVIK og klikk på denne. Merk område vi skal regne ut gjennomsnittsavviket for og klikk ok. Det siste spredningsmålet vi skal se på, er det vi kaller variasjonsbredde. Variasjonsbredde er nok det spredningsmålet som brukes mest i grunnskolen. Variasjonsbredde er definert som differansen mellom største verdi og minste verdi. Det er ingen egen funksjon i Excel som beregner variasjonsbredden, så vi må lage en funksjon selv. Det er imidlertid ganske enkelt. Vi har en funksjon som heter STØRST som finner største verdi i datasettet, og en funksjon som heter MIN som finner minste verdi i datasettet. For å finne største verdi blant gjestene på destinasjon A flytter du først musen til rute C23. Let deretter frem funksjonen STØRST, og merk området du vil søke gjennom, i vårt tilfelle A5 til D9. Funksjonen MIN fungerer på samme måte. Når vi nå skal finne variasjonsbredden, beregner vi den ved å trekke største verdi fra minste verdi. I vårt tilfelle blir det =C23-C24 og tilsvarende =I23-I24 for gjestene fra destinasjon B. Regnearket ditt bør nå se ut omtrent som det som er vist i innledningen på denne øvelsen.

Oppgave

a) Det er stor forskjell både på standardavvik, gjennomsnittsavvik og variasjonsbredde mellom gjestene på destinasjon A og B. Hva er årsaken til dette?

b) Se kun på gjennomsnittet og standardavviket eller gjennomsnittlig avvik for destinasjon A og B. Hvis du tar utgangspunkt i disse dataene, hvor vil du anbefale en ungkar i begynnelsen av 20-årene å reise? Hvor vil du råde en familie med unger å reise? Begrunn.

c) Hvorfor er det viktig at en tar med både sentralmål og spredningsmål i denne type datasett? Grunngi svaret.

## Øvelse 5: Regresjon

Denne øvelsen ligger litt i grenseland mellom statistikk og funksjonslære. Vi har likevel tatt den med under statistikk da temaet vanligvis behandles under statistikk i de fleste bøker. Vi skal se på 3 ulike problemstillinger i denne øvelsen. Først skal vi se på sammenhengen mellom alder og makspuls til 16 tilfeldige personer. Deretter skal vi se på hvordan folketallet i Norge har endret seg fra 1990 og frem mot i dag (2006). Vi skal også lage prognoser for utviklingen i tiden fremover. Til slutt skal vi se på hvordan bestanden av Sibirtigre har utviklet seg de seneste årene. I denne øvelsen skal vi også se litt på ulike typer regresjonsmodeller.

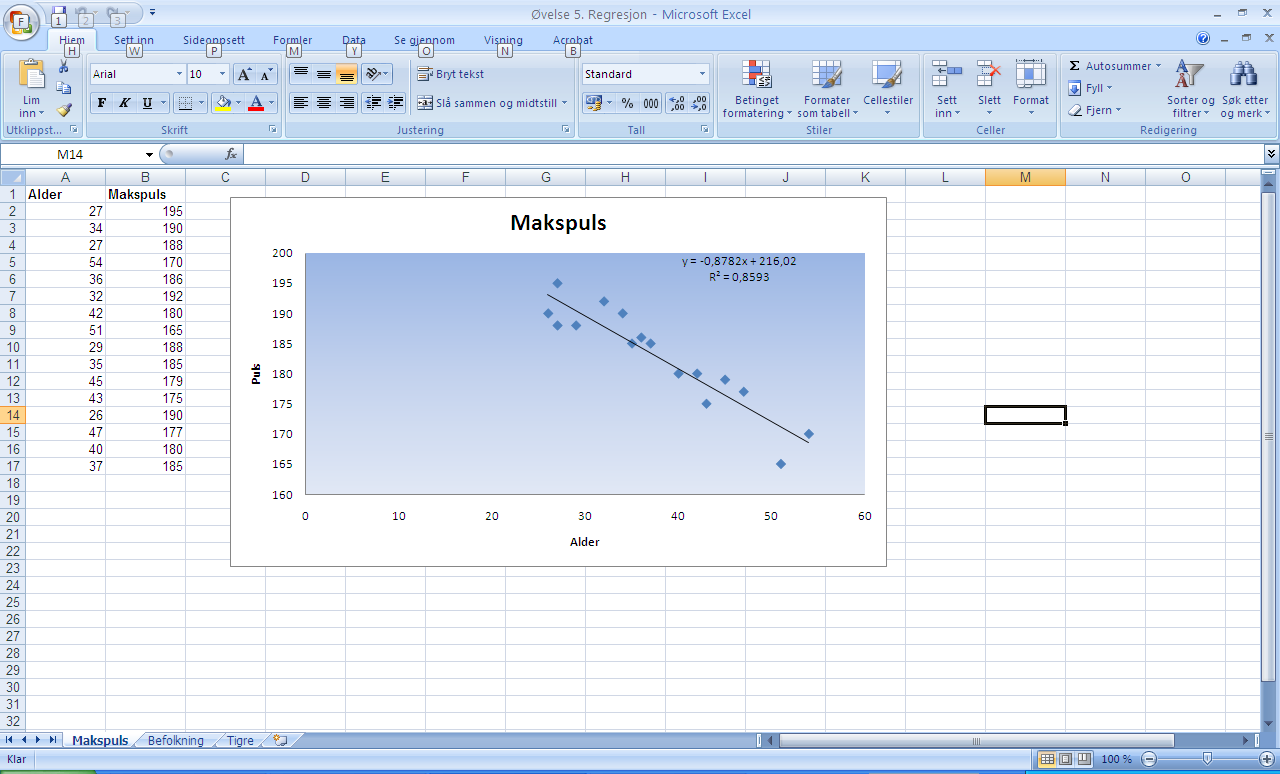
**Makspuls**

Fra den lokale sykkelklubben har vi plukket ut 16 tilfeldige personer. Vi har målt makspuls til alle, og i tabellen under ser du hva makspulsen og alderen er til disse personene.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Alder | 27 | 34 | 27 | 54 | 36 | 32 | 42 | 51 |
| Makspuls | 195 | 190 | 188 | 170 | 186 | 192 | 180 | 165 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Alder | 29 | 35 | 45 | 43 | 26 | 47 | 40 | 37 |
| Makspuls | 188 | 185 | 179 | 175 | 190 | 177 | 180 | 185 |

Vi er nå interessert i å kartlegge om det er det noe sammenheng mellom alder og makspuls. Vi skal først legge inn punktene i et Excel diagram og deretter få Excel til å hjelpe oss med å lage en rett linje som er tilpasset dataene. Vi skal med andre ord konstruere et regneark som vist under.



Start med å legge dataene inn i regnearket slik som vist på figuren på forrige side. Vi skal nå lage et punktdiagram for dataene. Merk A2 til B17 og klikk på Sett inn og velg deretter Punkt. Vi får da opp et diagram med punktene våre. Det kan variere litt hva du får opp som enheter på y-aksen. Vi skal låse disse verdiene. Høyreklikk på et av tallene på y-aksen og velg Formater akse. Velg så Skala. Sett minimum til 150, og ta eventuelt vekk merket som står før minimum. Da blir skalaen på y-aksen låst, og 150 blir minste verdi. Hvis du ønsker å ta bort støttelinjene, høyreklikk på en av linjene og klikk så på fjern. Plottet du har fått opp ser nå antagelig ut omtrent som dette:

Ser det ut som det er noe sammenheng mellom alder og hvilepuls? Excel kan nå legge inn en linje som er tilpasset dataene. Det mest vanlige er å bruke en rett linje, men også andre typer kurver kan brukes, som for eksempel andregradsfunksjoner, eksponentialfunksjoner etc. Excel beregner linjene etter en bestemt metode som kalles minste kvadraters metode. Den går ut på at summen av avstandene fra hvert punkt til linjen opphøyd i annen skal være minst mulig. Dette er meget arbeidskrevende å regne ut for hånd, selv med kalkulator. Excel ordener imidlertid dette på et blunk. Vi skal se på hvordan det kan gjøres. Start med å høyreklikke på et av punktene. Velg så Legg til trendlinje. Velg her lineær.. Hak ut rutene foran teksten Vis formel i diagrammet og hak også ut ruten foran teksten Vis R-kvadrat verdi i diagrammet. Vi har nå fått lagt til linjen, og Excel har også beregnet likningen for linjen. Vi ser vi har fått frem at . Dette tallet sier oss noe om hvor nært punktene ligger linjen. Verdien til  vil alltid ligge mellom -1 og 1. Den vil være negativ når linjen synker som i vårt tilfelle. Jo nærmere  er -1 eller 1 jo bedre samsvar er det mellom linje og punkter. Hvis , ligger punktene svært spredd i forhold til linjen. I vårt eksempel får vi . Når verdien ligger så nært -1, regner vi det som meget godt samsvar.

Oppgave

a) Hvordan samsvarer linjen med punktene? Er det mye avvik eller ligger det innenfor det akseptable?

b) Bruk modellen til å gi et anslag på hva hvilepulsen vil være for en 60 åring og for en 80-åring.

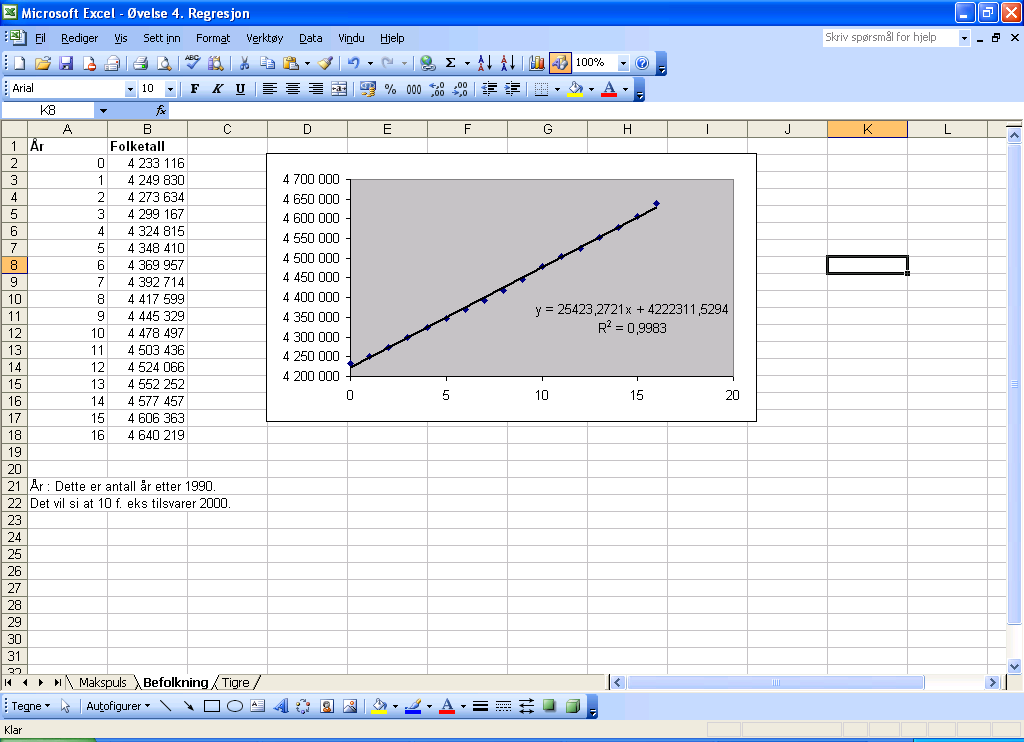
c) En tommelfingerregel sier at en persons makspuls vil være 220 minus alderen. Hvordan samsvarer vår modell med denne tommelfingerregelen?

**Befolkningsutvikling i Norge**

Vi skal nå se på et annet eksempel der vi tar for oss folketallet i Norge fra 1990 og frem til i dag (2006). I tabellen under ser du en oversikt over hvor stor befolkningen har vært i Norge fra 1990 og frem til 2006. Dataene er hentet fra den elektroniske utgaven av statistisk årbok, som du finner på adressen <http://www.ssb.no/aarbok/>

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **År** | **Folketall** |  | **År** | **Folketall** |
| 1990 | 4 233 116 |  | 1999 | 4 445 329 |
| 1991 | 4 249 830 |  | 2000 | 4 478 497 |
| 1992 | 4 273 634 |  | 2001 | 4 503 436 |
| 1993 | 4 299 167 |  | 2002 | 4 524 066 |
| 1994 | 4 324 815 |  | 2003 | 4 552 252 |
| 1995 | 4 348 410 |  | 2004 | 4 577 457 |
| 1996 | 4 369 957 |  | 2005 | 4 606 363 |
| 1997 | 4 392 714 |  | 2006 | 4 640 219 |
| 1998 | 4 417 599 |  |  |  |

Vi skal konstruere et regneark etter samme mal som i forrige eksempel. Du skal ta utgangspunkt i dataene som er gitt i tabellen over og bruke samme fremgangsmåte som i forrige eksempel til å lage et regneark som viser befolkningsutviklingen fra 1990 og frem til 2006. Det kan være lurt å la år 1990 tilsvare 0, år 1991 tilsvare 1 osv. Regnearket bør se ut omtrent som vist på neste side.



Oppgave

a) Hvordan samsvarer linjen som er tegnet opp med dataene våre?

b) Vi kan bruke modellen vår til å gi en prognose på hva folketallet blir i fremtiden. Bruk funksjonen som Excel har beregnet til å gi et anslag på hva folketallet vil være i år 2020 og i år 2050. Hvilke forutsetninger må du gjøre for å kunne bruke en slik modell?

c) Regnearket kan også tilpasses slik at linjen tegnes opp for de neste årene. Dette kan gjøres ved å høyreklikke en plass på trendlinjen. Velg deretter Formater trendlinje. Klikk på Alternativer. Velg så Forover og velg antall år frem i tid du vil at linjen skal tegnes. Prøv med litt forskjellige verdier og se hvordan diagrammet ser ut.

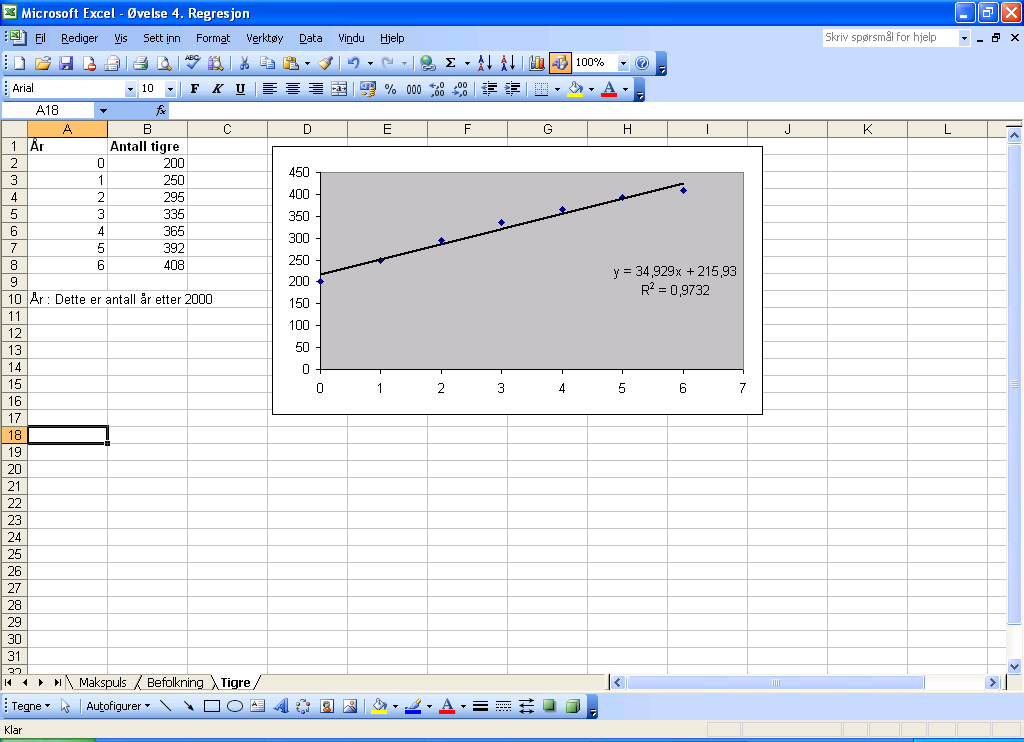
d) På websidene til statistisk sentralbyrå ligger det prognoser for befolkningsutviklingen. Gå inn på <http://www.ssb.no/aarbok/> og velg deretter befolkning. Nederst på siden ligger det en side med prognoser. Gå inn på denne og sammenlikn tallene med det du har regnet ut i spørsmål b). Hvordan samsvarer tallene? Hva kan årsaken være til eventuelle forskjeller?

**Bestanden av sibirtigere**

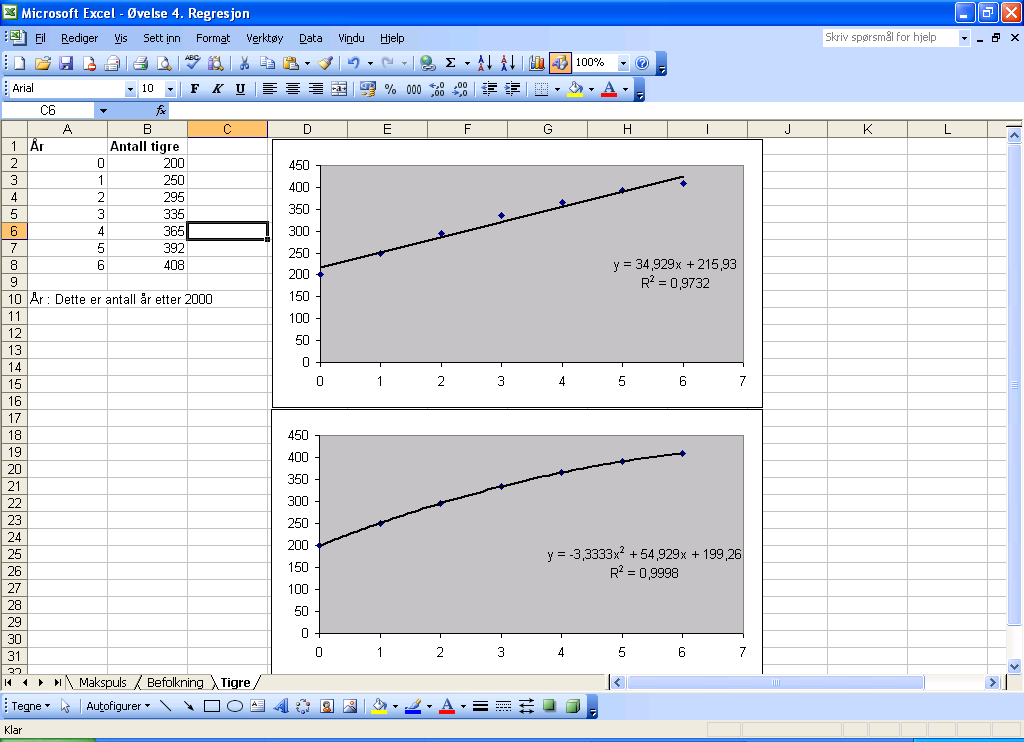
Vi skal til slutt i denne øvelsen se på et eksempel der vi ser på bestanden av Sibirtigere. I de to foregående øvelsene har vi brukt en lineær modell. Det vil si vi har brukt en rett linje tilpasset til dataene. Det skal vi gjøre her også, men i tillegg skal vi se på hvordan en kvadratisk funksjon kan brukes. Vi tenker oss at vi har hatt følgende utvikling i bestanden fra år 2000 og frem til 2006.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| År | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 |
| Antall tigere | 200 | 250 | 295 | 335 | 365 | 392 | 408 |

Vi skal starte med å lage et regneark etter samme mal som de to foregående oppgavene. Vi skal da få opp et regneark omtrent som følger:



Vi ser her at den lineære modellen er rimelig bra. Dataene ligger ganske nær den rette linjen som er trukket opp. Vi skal nå tilpasse en andregradsfunksjon til dataene. Det kan gjøres på flere måter. Vi kan lage et nytt diagram i det samme regnearket. Vi merker dataene og lager et nytt punktdiagram, og så legger vi til en trendlinje. Du skal klikke på polynom når du kommer til dialogboksen etter du har valgt Legg til trendlinje. Sørg for at orden er satt til 2. Gå så til Alternativer. Hak ut rutene foran teksten Vis formel i diagrammet, og hak også ut ruten foran teksten Vis R-kvadrat verdi i diagrammet. Klikk på ok. Du bør nå ha fått frem et regneark omtrent som vist under.



Oppgave

a) Hvilken modell synes du er den som beskriver virkeligheten best?

b) Bruk begge modellene til å finne et anslag for hvor mange tigere det vil være i år 2010 og 2020.

c) Drøft kort hvilken modell du tror vil være best egnet i fremtiden.

d) Prøv litt med andre typer funksjoner enn den lineære funksjonen og andregradsfunksjonen. Er det noen av de andre du synes passer bedre?