OPPGAVEHEFTE FOR STK1000 – KAPITTEL 1

OBS: Før du starter, les følgende viktige beskjed: Siden det gis ganske (og noen ganger svært) detaljerte Minitab-beskrivelser er det ikke noe poeng i seg selv å bare klippe og lime inn Minitab-outputtet. Poenget er å ha fornuftige og gjennomtenkte kommentarer til outputtet Minitab gir og ikke vise at man klarer å følge en oppskrift!

1. Oppgaver fra Kapittel 1

Regneoppgaver. Til tross for at alle har tilgang til en datamaskin som kan beregne oppsummerende statistikk fra et tallmateriale skal man ha gjort dette for hånd minst en gang for å "få følelsen" av hva som skjer. Disse oppsummerende metodene følger oss gjennom alle deler av kurset som grunnstenene til de statistiske verktøyene vi skal lære å bruke. Gjør derfor følgende oppgaver for hånd.

Ta det dessuten med ro; regneoppgavene for de kommende ukene går ikke ut på å plusse sammen mange tall eller andre "barneskoleaktige" oppgaver, og dette blir den eneste uken hvor dere settes til slikt.

Oppgave 1. Først en oppvarmingsoppgave for å bli kjent med summasjonstegnet

$$\sum,$$

som er den greske bokstaven Sigma. Hvis du vet hvordan dette tegnet fungerer kan du hoppe over denne oppgaven (bortsett fra a!).

Følgende tabell er åtte tilfeldige tall angitt med *to gjeldende siffer*.

2.53	1.37	-1.51	0.23	0.34	-2.21	0.00	1.10

(A) Den nest siste observasjonen er 0.00, men er allikevel ikke nødvendigvis nøyaktig null. Innenfor hvilket intervall[1] vet man at dette tallet er?

(B) Kall den første observasjonen x_1, den andre observasjonen x_2, osv., opp til at den åttende kalles x_8. Vi har at

$$\sum_{i=1}^{8} x_i = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8.$$

Vi har at $\sum_{i=1}^{8} x_i$ leses som "summen av x_i fra i lik 1 til 8". Betydningen overføres helt likt til flere tall, og man har altså

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_{n-1} + x_n$$

hvis man har n tall, der n hos oss er 8. Ofte skrives dette uten summasjonsgrenser, og man oppfatter dette dermed som "summer alle x'ene". Altså,

$$\sum x_i = \sum_{i=1}^{n} x_i.$$

Date: 1. september 2009.

[1] Et intervall er alle tallene mellom to grenser. F.eks er intervallet $[0, 1]$ alle tallene fra 0 til 1.

Regn derfor ut

$$\sum_{i=1}^{8} x_i$$

for hånd eller på kalkulatoren hvis du heller vil det.

(C) Når man ikke har x_i innenfor summetegnet, men f.eks x_i^2, eller en annen operasjon som avhenger av x_i, betyr det at man utfører denne operasjonen (dvs, f.eks å opphøye tallet i annen) på alle tallene for så å summere disse. For eksempel er

$$\sum_{i=1}^{8} x_i^2 = x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 + x_6^2 + x_7^2 + x_8^2.$$

Regn ut $\sum_{i=1}^{8} x_i^2$ og vis at det blir $16.8205 \approx 16.8$.

(D) Via litt algebra kan man vise at det empiriske standardavviket

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

også kan regnes ut via

$$\sqrt{\frac{n(\sum x_i^2) - (\sum x_i)^2}{n(n-1)}}.$$

Dette skal dere ikke vise, men denne formen er noe lettere å bruke når man skal regne for hånd (eller via en kalkulator som ikke gjør slikt automatisk). Regn ut empirisk standardavvik for datasettet over. Svaret er 1.530298, men siden vi bare har oppgitt observasjonene i to gjeldende siffer skriver vi også svaret med to gjeldende siffer, som vil si 1.53.

Oppgave 2. Et viktig diagnostisk hjelpemiddel består i telling av røde blodlegemer. Nedenfor er resultatet gitt av en rekke slike tellinger på blodprøver fra en bestemt person. Mengden blod er den samme ved hver telling.

Antall røde blodlegemer ved 25 tellinger (angitt i ordnet rekkefølge):

14	17	18	19	21	22	23	24	25	26
26	26	26	27	28	28	29	31	31	32
32	32	32	33	34					

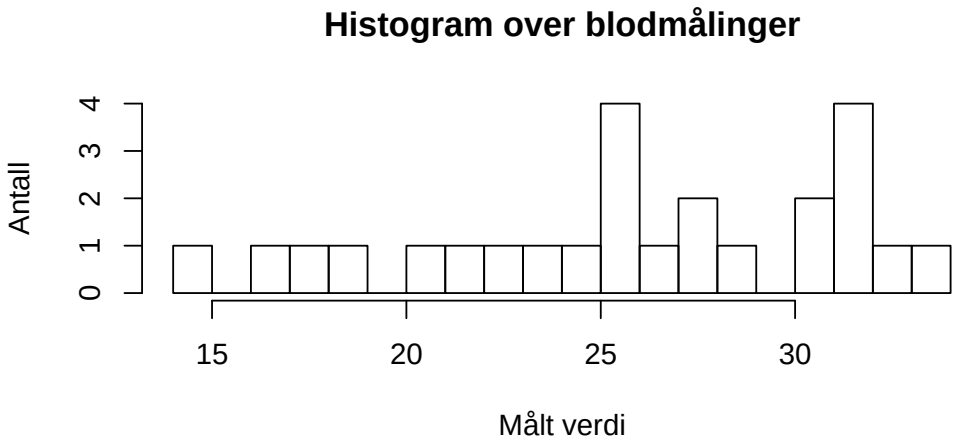
(A) Tegn et histogram over fordelingen av tallene ved å velge en passende søylebredde.

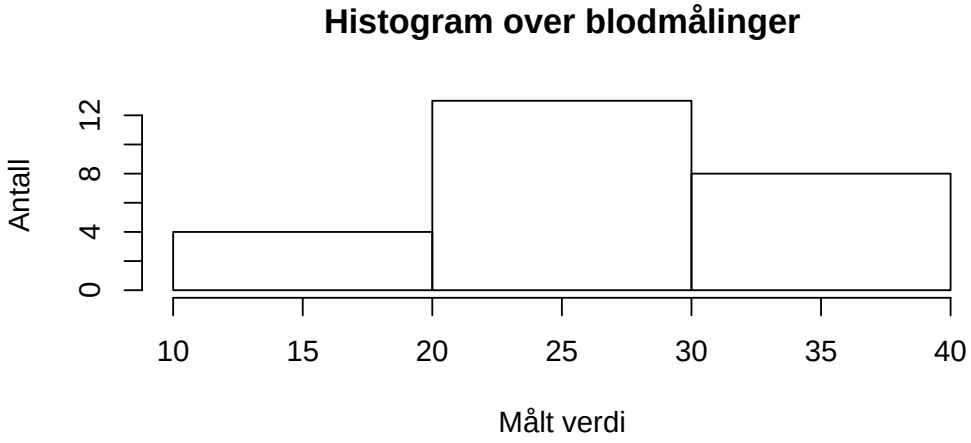
(B) Jo flere datapunkter man har, jo mindre søylebredde kan man tillate seg og fortsatt få et fornuftig inntrykk av datasettet. I figuren nedenfor har man valgt å lage et histogram med 17 og 3 søyler. Kommenter hvorfor dette er lite hensiktsmessige valg.

(C) Regn ut gjennomsnitt og standardavvik. Bruk metoden fra oppgave 1 b for å regne ut standardavviket.

(D) Finn første, andre og tredje kvartil, dvs 25 %, 50 % (medianen!) og 75 % kvantilene. Finn også interkvantilavstanden.

(E) Tegn så et boksplott og sammenlikn hva slags inntrykk du får av datasettet med det du fikk av histogrammet.





(F) Datasettet er ordnet i stigende rekkefølge, og man har altså ikke tatt vare på rekkefølgen av observasjonene (som er for å være ærlig gjort for at dere skal slippe å sortere observasjonene for hånd i oppgave *(E)*, men la oss for diskusjonens skyld si at man fikk datasettet fra et laberatorie i en slik form). Virker dette rimelig, eller burde man også ha tatt vare på denne rekkefølgen? Kan du tenke ut en situasjon der det kunne være nyttig å ha med rekkefølgen?

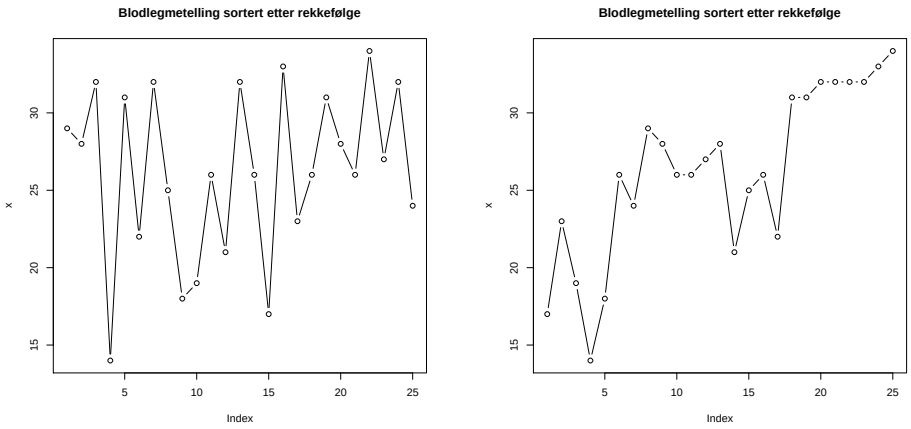
La oss anta at tellingen gjøres av mennesker og ikke maskinelt. Ville du som ansatt statistiker reagert likt hvis du fikk de to datasettene under? Disse består *tallmessig* av de samme observasjonene, og eneste forskjell er rekkefølgen på observasjonene.

29	28	32	14	31	22	32	25	18	19
26	21	32	26	17	33	23	26	31	28
26	34	27	32	24					

TABELL 1. Datasett 1

17	23	19	14	18	26	24	29	28	26
26	27	28	21	25	26	22	31	31	32
32	32	32	33	34					

TABELL 2. Datasett 2



(a) Plott av datasett 1

(b) Plott av datasett 2

FIGUR 1. Plott av de to datasettene.

Minitab-oppgaver.

Oppgave 3. Oppgave 2 i forrige seksjon skulle gjøres for hånd. Vi skal nå gjøre den igjen (bortsett fra første spørsmål), men denne gangen skal vi bruke Minitab. Oppgaven er ment for å bli kjent med Minitab samt hvordan man tar vare på Minitabutskrifter i Word (eller tilsvarende officepakke som Open Office), og har derfor mer detaljerte Minitab-hint enn kommende oppgaver.

Start først opp Minitab og Word, og last inn filen "blod.txt" i Minitab. Dette er altså ikke en fil i "Minitab-formatet", men en ren tekstfil. Siden det er vanlig å spre datafiler i ren tekst må man også vite hvordan man legger inn data fra disse[2]:

i. Åpne filen i Notepad(eller tilsvarende tekstredigeringsprogram, som Word)
ii. Marker all tekst, enten ved hjelp av musen eller ved å trykke CTRL samtidig som a[3].
iii. I rullegardin menyen, velg **Rediger** og så **Kopier**, eller trykk CTRL samtidig som c[4].
iv. Gå tilbake til Minitab. Trykk på ruten C1-1, dvs den øverste *hvite* ruten til venstre, med venstre museknapp.
v. I rullegardinmenyen, velg **Edit** og så **Paste Cells**, eventuelt trykk CTRL samtidig som v.

[2]Denne forklaringen kan så klart hoppes over hvis du kan litt om data.
[3]For "marker **A**ll".
[4]For "**C**opy", også kjent som kopier.

Vi har nå lagt inn selve datasettet. Den grå øverste raden i Minitabs regneark er *satt av til navn på kolonner med data*. Datasettet vårt er tellinger av røde blodlegemer. Navngi derfor hele C1 kolonnen "Røde blodlegemer". Det er fornuftig å *alltid* navngi datakolonner, så man ikke mister oversikten, og så man slipper å huske på hvilket kolonnenummer som hører til hva.

Minitab-hint for selve oppgavene:

(B) Tegn et histogram av datasettet.

Histogram tegnes via rullegardinsmenyvalg gitt ved

Graph $\mapsto$ Histogram.

Velg "Simple" og klikk OK, og dobbeltklikk på "Røde blodlegemer". Disse havner dermed på "Graph variables"[5]. Klikk så OK. Opp kommer så Minitabs autogenererte histogram.

Vi vil nå lagre dette i Word. Klikk derfor hvor som helst inni histogramvinduet og trykk CTRL samtidig som c. Dette kopierer bildet av histogrammet til minnet på maskinen. Gå så inn i Word, skriv en overskrift (som "Oppgave 3") og lim inn bildet ved å trykke CTRL samtidig som v.

(D) Regn ut gjennomsnitt og standardavvik.

Minitab-hint: Lukk histogramvinduet som du har tatt vare på i Word. Gå så på

Stat $\mapsto$ Basic Statistics $\mapsto$ Display Descriptive Statistics.

Velg at vi er interessert i "Røde blodlegemer" ved å dobbeltklikke på dens tekstfelt og trykk OK.

I "Sessions"-tekstfeltet får vi nå en del oppsummerende statistikk, som *inkluderer* de to tallene vi er interessert i, nemlig "mean" som altså er gjennomsnittet og "Standard Deviation" (forkortet til "StDev") som er standardavviket[6].

Vi vil nå lagre dette i Word. Det er to måter å gjøre dette på;

i. Marker all den teksten Minitab gav ut for så å kopiere og lime det inn i Word. Dette er mer informasjon enn man trenger, og ikke særlig lettlest eller oversiktlig.
ii. Finn akkurat de tallene du søker, og skriv en vanlig norsk setning i Word om at vi fant at gjennomsnitt var det og det, samt standardavviket var det og det.

(E) Finn kvantiler og interkvantilavstanden.

Kvantiler har vi allerede funnet i forrige punkts Minitabutskrift. Selve Interkvantilbredden kunne vi spesifikt ha bedt Minitab om å regne ut i tillegg, som beskrevet i en fotnote på forrige oppgave, men regn det heller ut for hånd. Dvs, under rullegardinsmenyen, velg

Tools $\mapsto$ Microsoft Calculator

og ta Q3 - Q1. Rapporter resultatene i Word. Det kan være hensiktsmessig å skrive opp disse oppsummerende tallene sammen, og å ikke dele opp besvarelsen i punkter (D) og (E).

---

[5] Alternativt kunne man ha klikket på tekstfeltet til "Graph variables", og skrevet inn "C1".

[6] Hvis vi ikke var interessert i alle de andre tallene, eller hvis vi var interessert i andre oppsummerende tall, kunne vi ha klikket på "Statistics" i "Display descriptive statistics"–vinduet for så å spesifisert nærmere hva vi ville Minitab skulle gi ut.

(F) Tegn et boksplott av datasettet.

I Minitab gjøres dette via

Graph $\mapsto$ Boxplot.

Velg "Simple" under "One Y". Velg så at du vil tegne opp "Røde blodlegemer". Ta vare på dette bildet i Word ved å lime det inn. Men det er ikke hensiktsmessig å lagre dette bildet i den størrelsen det automatisk kommer i. Klikk derfor på det innlimte boksplottet i Word. Da kommer det noen små punkter i kantene på bildet. Dra disse innover for å skalere bildet til en passende størrelse. Skaler også histogrammet så det passer til resten.

Sett til slutt på en fornuftig og oppsummerende kommentar til oppgaven.

Oppgave 4. Både histogrammer og boksplott er gode og ofte brukte verktøy som gir en rask oversikt over mange tall på en gang. Spesielt for svært store datasett er vi avhengig av deskriptive (beskrivende) statistiske metoder for å "se" et tallmateriale.

Last inn datasettet "families.mtw" i Minitab. Dette datasettet er en *stor* tabell med informasjon om 43886 familier i en storby i USA. Informasjonen som er lagret er beskrevet i tabellen nedenfor.

Variabelnavn	Forklaring
"Type"	1 = Mann og kone, 2 = Alenefar, 3 = Alenemor
"Persons"	Antall personer i familien
"Children"	Antall barn i familien
"Income"	Inntekt i dollar
"Region"	0 = Nord, 1 = Øst, 3 = Sør, 4 = Vest
"Education"	Tallindikator på utdannelse fra 31 (mist utdannelse) til 46 (PhD)

TABELL 3. Forklaring til "families.mtw"

(A) Beregn oppsummerende statistikk ved hjelp av Minitab som i den forrige oppgaven. Dvs, gå på

Stat $\mapsto$ Basic Statistics $\mapsto$ Display Descriptive Statistics.

Lag deretter et histogram over inntekt, men spesifiser at Y-aksen skal måles i prosent og ikke antall.

Dette gjøres ved å gå inn på **Scale $\mapsto$ Y-scale type** rett før man trykker "OK" for å lage histogrammet. Der spesifiserer man "Percent" istedenfor "Frequency".

(B) Hvilket intervall vil du si at beskriver familien til "Average-Joe's" inntekt best?

(C) Beskriv fordelingen med stikkordene unimodal eller multimodal, symmetrisk eller høyre- og venstreskjev. Forklar hva beskrivelsen din av histogrammet betyr i dagligdags språk. Ser det f.eks ut til å være strengt skilte klasser i befolkningen? Og hvordan ville histogrammet isåfall sett ut?

(D) "Mye" og "lite" defineres gjerne utifra hva som er vanlig. Hva vil du si er å tjene mye og hva vil du si er å tjene lite i forhold til denne befolkningen?

(E) Lag et nytt histogram for inntekt, men gå på **Scale** $\mapsto$ **Y-Scale Type** og velg "Accumulate values accross bins". Dette histogrammet viser prosentandel[7] av observasjonene som er *mindre eller lik* markert X-verdi. Hvor mange prosent tjener mye, og hvor mange prosent tjener *veldig* mye? Og hvor mange prosent tjener lite, og hvor mange prosent tjener *veldig* lite?

Oppgave 5. Boksplott kan blant annet brukes til å visualisere sammenhenger i et datamateriale. Vi skal her se på hva slags tentative konklusjoner man kan få ut av inntektsdatasettet brukt i oppgave 4 ved hjelp av boksplottet.

(A) Lag et boksplott av inntekt for hver type utdannelse via

Graph $\mapsto$ **Boxplot** $\mapsto$ **With Groups**

og spesifiser INCOME som "Graph variables" med EDUCATIO som "Categorical variables for grouping".

(B) Hva betyr alle stjernene? Boken har en diskusjon om at det kan være noe feil med "uteliggende" observasjoner[8]. Under antagelse om at ingen rike unngår å rapportere alt de har tjent, er det snakk om at det her er noe "feil" med disse observasjonene? Hvordan kan disse tolkes?

(C) Og hva er strekene som strekker seg ut av boksene? Til tross for at man tar en høy utdannelse, ser man at det er noen som ender opp med nær null eller negativ inntekt. Hvordan ser man dette?

(D) Vil du si at det er en trend i inntektsfordelingen i forhold til utdanningsnivå, i hvert fall for medianene? Det siste nivået er å ta en doktorgrad, mens den nest høyeste representerer en profesjonsutdannelse. Lønner det seg å ta en doktorgrad? Er interkvantilbredden den samme for alle grupper? Hvor er den minst og størst?

Oppgave 6. Gjør oppgave 1.8 i boken. Datasettet ligger lagret som "oppgave1.8.mtp". Minitab-hint:

(A) Minitab kan gjøre dette via

Calc $\mapsto$ **Column Statistics** .

Det automatiske valget er "Sum", som er det vi vil ha. Altså må vi bare velge hvilken kolonne som skal summeres. Velg "Weight" i "Input variable", og la "Store result in" være tomt. Resultatet kommer opp på "Sessions"-vinduet.

(B) Gå på

Graph $\mapsto$ **Bar Chart** ,

og velg "Values from a table" på "Bar represents:", og trykk ok. Velg "Weight" i "Graph variable", og "Material" i "Categorical variables".

(C) Gå på

Graph $\mapsto$ **Pie Chart** .

Velg "Weight" i "Summary variables", og "Material" i "Categorical variable".

Oppgave 7. Gjør oppgave 1.27 i boken. Datasettet finnes på "ta01_007.MTP". Minitab-hint: Gå på

Graph $\mapsto$ **Time Series Plot** .

[7] Dvs, den viser prosentandel siden du har stilt inn dette før i oppgaven.

[8] Det vil si "Outlying observations" på engelsk

Velg "Multiple" under menyen som nå kommer og trykk på OK. Legg inn "C2 pasadena" og "C3 reading" under "Series". For å få Minitab til å skjønne tidsskalaen, gå så på "Time/Scale". Under "Time Scale", velg "Stamp", og klikk inn "C1 year" på "Stamp Columns".

Oppgave 8. Gjør oppgave 1.54 i boken. Datasettet finnes på "ex01_054.MTP". For å finne oppsummerende statistikk for "gpa" og "iq" samtidig, kan man gå på

Stat $\mapsto$ **Basic Statistics** $\mapsto$ **Display Basic Statistics**

og sett inn både "gpa" og "iq" i "Variables". Dessuten; Minitab kan lage et stem-and-leaf-plot ved

Graph $\mapsto$ **Stem-and-leaf .**

Oppgave 9. I denne oppgaven skal vi se nærmere på fordelingen av fødselsvektene til et tilfeldig utvalg av 100 jenter født i Norge i 1985 der svangerskapet varte mellom 37 og 43 uker. Dataene ligger under "FVEKT.MTW".

(A) Lag et histogram over dataene (**Graph** $\mapsto$ **Histogram**).

(B) Lag også et histogram der det er tegnet inn en tilpasset normalfordeling i samme figur via

Stat $\mapsto$ **Basic Statistics** $\mapsto$ **Display Descriptive Statistics**.

Velg "Graphs" og kryss av for "Histogram of data, with normal curve".

(C) Lag et normalfordelingsplott av datene (**Graph** $\mapsto$ **Probability plot**. Se eventuelt under "Options" hvordan du kan fjerne de ekstra linjene fra plottet). Merk at normalfordelingsplottet i MINITAB bytter om x- og y-aksene i forhold til det som står i læreboka (men normalfordeling svarer fortsatt til at plottet er noenlunde rettlinjet). Syns du det ser ut som om en normalfordeling beskriver fordelingen av fødselsvektene rimelig godt?

OBS: Spørsmålet over er stilt som "syns du det ser ut som ..." og er veldig kvalitativt. Kvalitativ synsing burde være på en ordentlig bakgrunn, og man burde ha en viss erfaring med normalfordelingsplott for å kunne synse med sikkerhet. Denne erfaringen får man etter å ha sett hva slags rammer et normalfordelingsplott burde ha, og vi skal tilegne oss dette i en av obligoppgavene hvor vi også skal besøke disse 1985-babyene. Det sentrale spørsmålet er: Hvis vi har mange grupper med normalfordelte observasjoner, hvordan vil normalfordelingsplottene til hver gruppe avvike fra en rett linje? Hva er det verste man kan akseptere som OK, og hva slags konklusjoner kan man komme frem til hvis man sier at normalfordelingsplottet er OK? Betyr det at datasettet *er* normalfordelt? Svaret på det siste er "nei, men at det ofte ikke har så mye å si". I oblig 1 skal vi se at svaret er "nei". Mens det mer sofistikerte tilegget "men at det ofte ikke har så mye å si" kommer vi først til under diskusjoner av robusthet i del to av boken. Dette tilegget er en av to gode grunner som gjør at normalfordelingen er *verdens viktigste statistiske fordeling* (den andre er den såkalte "sentralgrense effekten" som mange har hørt om fra før fra populærvidenskapelige kilder, men som også er ofte misforstått. Vi skal møte denne effekten først i kapittel fem!)

(D) Beregn gjennomsnittlig fødselsvekt og (empirisk) standardavvik for fødselsvektene via

Stat $\mapsto$ **Basic Statistics** $\mapsto$ **Display Descriptive Statistics**.

Undersøk hvor godt "68 - 95 - 99.7 regelen" (side 71 i læreboka) passer for dataene. For å gjøre dette er det hensiktsmessig å sortere dataene i stigende rekkefølge (**Data** $\mapsto$ **Sort**).

Oppgave 10. I sykdommen malaria kommer parasitter inn i de røde blodlegemene og kan ødelegge dem. De infiserte blodlegemene vil ofte være forstørrede. Det viser seg at diameter (målt som maksimal diameter, da legemene ikke er sirkelformede) for infiserte blodlegemer tilnærmet følger en normalfordeling med forventning 9.6 og standardavvik 1.0 mikron. Normale blodlegemer følger en normalfordeling med forventning 7.6 og standardavvik 0.9 mikron.

OBS: Etter oppgaveteksten følger en beskrivelse for hvordan man kan få Minitab til å gjøre ønskede operasjoner. Husk å lim inn resultatene i en Word-fil (eller tilsvarende program) og skriv en "mikrorapport".

(A) Tegn de to normalfordelingene i samme graf. Du vil se at de overlapper – hvilke problemer skaper dette for identifiseringen av infiserte blodlegemer på bakgrunn av størrelse?

(B) Hvilken andel av ikke-infiserte blodlegemer fra en pasient vil ha diameter større enn 9.6 mikron? Hvilken andel av infiserte blodlegemer fra en pasient vil ha diameter mindre enn 7.6 micron?

(C) Hvor stor andel av de infiserte blodlegemene vil ha en diameter mindre enn $7.6 + 2 \times 0.9$ mikron? Hvor stor andel av de ikke-infiserte blodlegemene fra pasienten vil ha en diameter større enn $9.6 - 2 \times 1.0$ mikron?

Dette løses i Minitab som følgende:

- Lag først en vektor med tall som skal brukes til plotting (x-verdier) av normalfordelingene:

  Calc $\mapsto$ **Make Patterned Data** $\mapsto$ **Simple set of numbers**.

  Bruk følgende oppsett.

  ```
  Store patterned data in: C1

  From first value: 3

  To last value: 15

  In steps of: 0,025
  ```

- Gå så til

  Calc $\mapsto$ **Probability Distributions** $\mapsto$ **Normal**

  og klikk på **Probability density**.
- Legg inn Mean og Standard deviation, C1 som Input column, og lagre i C2.
- Gjør det samme for normalfordelingen for infiserte blodlegemer, men lagre i C3. Nå kan du gå til

  Graph $\mapsto$ **Scatterplot** $\mapsto$ **With groups**

  og legge inn C2 og C3 som Y, C1 som X i begge tilfeller. Huk av X-Y pairs form groups for å tegne de to normalfordelingene i samme graf.
- For å beregne andelene, klikk på

  Calc $\mapsto$ **Probability Distributions** $\mapsto$ **Normal**

og velg Cumulative probability i stedet for Probabillity density som over.