Projekt 8.9 Hvordan undersøges om et talmateriale normalfordelt?

Projektet drejer sig om at udvikle en metode, til at undersøge om et givet talmateriale med rimelighed kan siges at være normalfordelt. Det giver samtidig en teoretisk begrundelse for anvendelse af det såkaldte normalfordelingspapir, der er et praktisk værktøj til en hurtig undersøgelse af et datamateriale og til at give estimater for middelværdi og spredning. Projektet indeholder overlappende materiale ift. kapitel 8 i Hvad er matematik? 3 (HEM3) for at give en sammenhæng i projektet.

Normalfordelingen
Normalfordelingen er introduceret og grundigt behandlet i kapitel 8. Normalfordelingen præsenteres ofte som en "klokkeformet" kurve, men verden er fuld af forskellige klokkeformede kurver, så hvad er det lige, der udmærker den særlige kurve, vi kalder for en normalfordelingskurve, og som har fundet anvendelse overalt, hvor man anvender statistik.
Populært sagt kan man beskrive normalfordelingskurven som den kurve, der fremkommer, hvis man lader rigtig mange kugler falde ned gennem et meget stort *Galtonbræt*. Et Galtonbræt er præsenteret i bogens kapitel 8 og består af lange rækker af søm placeret med lige store mellemrum, hvor en kugle netop kan trille igennem. Rækkerne af søm er placeret forskudt for hinanden, så en kugle, der triller gennem en åbning falder ned på et søm der er placeret netop midt i mellemrummet. Når kuglen rammer dette søm vil den med lige stor sandsynlighed falde til venstre eller til højre. Når vi lader rigtig mange kugler trille igennem et øverste mellemrum, vil de - efter at have passeret rigtig mange rækker- lande i optællingsbokse med samme bredde som mellemrummene, og hvor således højderne angives af antallet af kugler, der netop landede der. Den kurve, som dette "pindediagram" tegner er en klokkeformet kurve, der tilnærmer en normalfordelingskurve bedre og bedre hvis vi vælger flere og flere kugler og flere og flere rækker, de skal passere.
En sådan bevægelse, hvor man med lige stor sandsynlighed kan gå til højre som til venstre, kaldes for en "random walk", og definitionen på en normalfordeling er derfor, at en størrelse (i statistik kalder vi dette for en *stokastisk variabel*) kaldes for normalfordelt, hvis den er fordelt på samme måde, som en random walk med et meget stort antal skridt. Meget stort betyder ideelt set uendeligt mange, og et forsøg, der tænkes gentaget uendeligt mange gange kaldes netop for et ideelt forsøg.

Når man lader antallet af kugler og antallet af rækker, der passeres, vokse mod uendelig, så vil der ske to ting:

- Grafen ville komme til at fylde uendeligt meget, hvis vi ikke skalerede ned. Det gør man løbende, så efter at have ladet n kugler trille ned skalerer vi slutværdierne på x-aksen ned med $\sqrt{n}$, hvorved spredningen hele tiden holdes på 1. Samtidig skaleres y-værdierne (højden af "pindene") ned med $\frac{\sqrt{n}}{2}$, hvilket sikrer, at det samlede areal under kurven hele tiden er 1. I grundbogen er der redegjort nærmere for dette.
- Grafen vil ændre sig fra at bestå af en samling af enkeltpunkter, vi forbinder, til at være en sammenhængende (kontinuert) kurve.

Det samlede areal kan måles ved samlede antal kugler, og derved ser man, at sandsynligheden for at en kugle lander mellem to værdier på x-aksen må svare til det relative antal kugler, der er landet mellem de to værdier. Men det betyder for den kontinuerte kurve i grænsen, at sandsynligheden for at ligge i et bestemt interval må svare til arealet under kurven mellem de to værdier, og dette beregnes som et integral.

Værdierne på x-aksen, hvorover vi har afsat højden af "pindene" ("kuglerne"), kaldes for *observationer.*

Den klokkeformede kurve der fremkommer på denne måde kaldes for *standardnormalfordelingskurven*. Den har middelværdi 0, idet man lægger x-aksen med begyndelsespunkt lige under det sted, hvor kuglerne trilles ned. Den gennemsnitlige afstand fra der, hvor en kugle lander, og ind til middelværdien kaldes for standardafvigelsen eller spredningen og denne er 1 for standardnormalfordelingen. Middelværdi betegnes med symbolet μ (græsk bogstav for m, udtales my) og spredningen betegnes med symbolet σ (græsk bogstav for s, udtales sigma). Man kan vise, at *standardnormalfordelingskurven* er graf for funktionen:

$$\varphi(x)=\frac{1}{\sqrt{2\cdot\pi}}\cdot e^{-\frac{1}{2}x^2}$$ 5

Normalfordelingen arver en række egenskaber fra random walk fordelingerne, fordi det er en grænseværdi for dem. Blandt de vigtigste er:

- ca. 68% af observationerne ligger i en afstand fra middelværdien på 1 spredning, dvs mellem $\mu-\sigma$ og $\mu+\sigma$.
- ca. 95% af observationerne ligger i en afstand fra middelværdien på 2 spredning, dvs mellem $\mu-2\sigma$ og $\mu+2\sigma$. Observationer, der ligger inden for denne afstand, kaldes for *normale*.
- ca. 0,25% af observationerne ligger i en afstand fra middelværdien, der større end 3 spredninger. Observationer, der ligger i denne afstand, kaldes for *exceptionelle*.

Øvelse 1

a) Tegn grafen for $\varphi(x)=\frac{1}{\sqrt{2\cdot\pi}}\cdot e^{-\frac{1}{2}x^2}$. Overvej selv indretningen af grafvinduet ud fra ovenstående.

b) Vis at arealet under kurven er 1.

Lad nu *Z* være en variabel, der angiver observationer, der er standardnormalfordelt, Dette skriver man ofte således:
$Z \sim N(0,1)$

Da arealet under kurven angiver sandsynligheder, så er fx $P(Z\le c)$ dels lig med arealet under kurven fra $-\infty$ til tallet *c* og samtidig lig med sandsynligheden for at en tilfældig observation er mindre end *c*:

$$P(Z\le c)=\int_{-\infty}^{c}\frac{1}{\sqrt{2\cdot\pi}}\cdot e^{-\frac{1}{2}x^2}\,dx=\frac{1}{\sqrt{2\cdot\pi}}\cdot\int_{-\infty}^{c}e^{-\frac{1}{2}x^2}\,dx$$

Indfører vi en ny variabel $Y=a\cdot Z+b$, så har vi:

$$P(Y\le t)=P(a\cdot Z+b\le t)=P\left(Z\le\frac{t-b}{a}\right)=\frac{1}{\sqrt{2\cdot\pi}}\cdot\int_{-\infty}^{\frac{t-b}{a}}e^{-\frac{1}{2}x^2}\,dx$$

Øvelse 2

Argumenter for, at *Y* må være en normalfordelt med middelværdi *b* og spredning a

Øvelse 3

Vis ved anvendelse af substitutionen $x=\frac{y-b}{a}$:

$$P(Y\le t)=\frac{1}{a}\cdot\frac{1}{\sqrt{2\cdot\pi}}\cdot\int_{-\infty}^{t}e^{-\frac{1}{2}\left(\frac{y-b}{a}\right)^2}\,dy=\int_{-\infty}^{t}\frac{1}{a}\cdot\frac{1}{\sqrt{2\cdot\pi}}\cdot e^{-\frac{1}{2}\left(\frac{y-b}{a}\right)^2}\,dy=\int_{-\infty}^{t}\frac{1}{a}\cdot\varphi\left(\frac{y-b}{a}\right)dy$$

Under integraltegnet står således tæthedsfunktionen for normalfordelingen *Y* med middelværdi *b* og spredning a. Betegnes middelværdi med μ og spredningen med σ, så gælder:

Tæthedsfunktionen for $Y \sim N(\mu,\sigma)$ er lig med $\frac{1}{\sigma}\cdot\varphi\left(\frac{y-\mu}{\sigma}\right)$

Øvelse 4

a) Opret i et værktøjsprogram grafen for $\varphi(x)=\frac{1}{\sqrt{2\cdot\pi}}\cdot e^{-\frac{1}{2}x^2}$. Bestem grafens vendepunkter og bestem afstanden fra disse og ind til 0.

b) Indfør skydere for middelværdi μ og spredningen σ og opret grafen for $\frac{1}{\sigma}\cdot\varphi\left(\frac{y-\mu}{\sigma}\right)$. Hold først spredningen fast på 1 og undersøg betydningen af μ. Hvor ligger grafens symmetriakse? Bestem grafens vendepunkter og bestem afstanden fra disse og ind til symmetriaksen.

c) Sæt middelværdien til 0 og varier spredningen. Hvor ligger grafens symmetriakse? Bestem grafens vendepunkter og bestem afstanden fra disse og ind til symmetriaksen.

d) Giv nu en samlet beskrivelse af det grafiske billede af tæthedsfunktionen $\frac{1}{\sigma}\cdot\varphi\left(\frac{y-\mu}{\sigma}\right)$

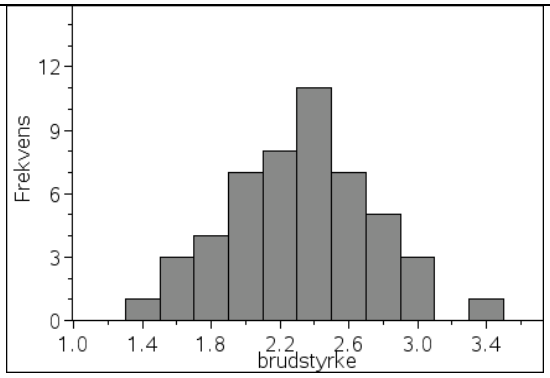
Anvendelse af histogram som tilnærmelse til en normalfordelingskurve

Når vi ønsker at undersøge om et givet talmateriale følger en normalfordeling, er det første naturlige skridt at repræsenterer talmaterialet med en graf, som kan sammenlignes med en normalfordelingskurve. For at illustrere dette tager vi udgangspunkt i det datasæt, der er behandlet i Grundbogen til B, kapitel 8, afsnit 3.4, og som handler om brudstyrker for hørgarn. Datasættet består af 50 målinger. Du kan hente en tabel over brudstyrkerne ***her.*** Gør det og gennemfør selv den følgende behandling af materialet.

1.73	2.86	2.94	2.63	2.40	2.79	2.32	2.77	2.55	1.89
1.99	1.69	2.50	2.30	2.47	2.23	2.31	2.52	1.98	2.60
2.65	1.40	2.07	2.92	2.12	1.63	2.44	2.03	3.02	2.71
2.13	2.74	1.73	2.20	1.95	2.35	1.92	2.39	2.26	1.52
2.36	2.37	2.12	2.64	3.30	2.40	1.78	2.02	2.15	2.16

Vi vil prøve at vurdere om brudstyrkerne med rimelighed kan siges at være normalfordelte. Vi grupperer datasættet i intervaller med samme intervalbredde. Her har vi valgt en bredde på 0,2. Vi tegner så et histogram over fordelingen og ser da noget der med lidt god vilje godt kunne ligne en klokkeformet fordeling. Men for at kunne sammenligne det med en normalfordeling må vi kende middelværdien, spredningen og arealet hørende til histogrammet. Da der er tale om en stikprøve bruger vi stikprøvespredningen:

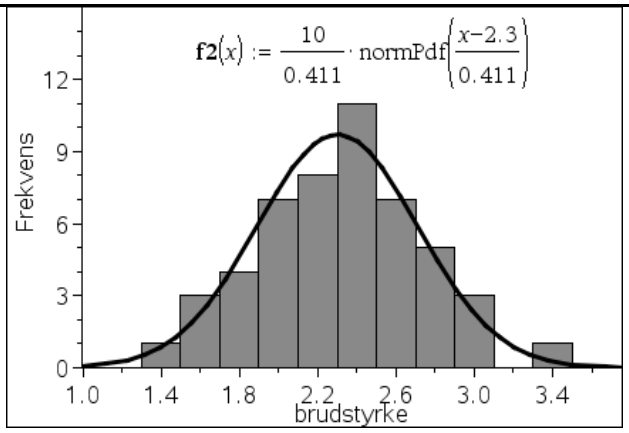
Titel	Statistik med én variabel
$\bar{x}$	2.299
$Sx := Sn_{-1}x$	0.410929
$\sigma x := \sigma_n x$	0.406798



Histogram tegnet med søjlebredden 0.2 og startværdi for yderste søjle 1.3.

Middelbrudstyrken er 2.30, mens stikprøvespredningen s er 0.411. Da der er 50 stykker hørgarn og grundlinjen i søjlerne er 0.2 fås et samlet areal på 10.

Vi tegner derfor fordelingskurven for normalfordelingen med den *samme middelværdi, spredning* og *areal*. Resultatet er igen rimeligt. Og så afhænger det i øvrigt af hvordan vi har opdelt histogrammet. Søjlen med 2.2 ligger lidt for lavt, mens søjlen med 2.4 ligger lidt for højt i forhold til det forventede, men ellers følger normalfordelingskurven histogrammet rimeligt pænt. Mange regneark kan konstruere det overlejerede normalfordelingsplot helt automatisk.



Histogram med overlejret normalfordelingskurve, der har den samme middelværdi, spredning og areal.

Transformation af data til undersøgelse af om datasættet er normalfordelt.
Som vi omtalte i indledningen er der mange klokkeformede kurver, og man vil ofte ønske noget mere overbevisende end ovenstående grafiske redegørelse. Og der findes faktisk en anden mulighed for at visuelt at vurdere om et givet observationsmateriale med rimelighed kan siges at følge en normalfordeling.

Hvor vi ovenfor ser på *histogrammer* og de tilhørende *tæthedsfunktioner* for normalfordelinger, så vil vi her se på *sumkurver* og de tilhørende *fordelingsfunktioner* for normalfordelinger. I sumkurver afsættes de kumulerede procenter over observationsintervallernes højre endepunkter. Det giver en kurve der har form som et langstrakt S og som forløber fra 0% til 100%. Den ideelle normalfordelingsfunktion har ingen øvre og nedre grænse for observationerne, så vi vil her aldrig nå helt ned til 0 eller helt op til 100%.

Ved *fordelingsfunktionen* hørende til en (stokastisk) variabel *X* forstår vi funktionen:

$$F(x) = P(X \leq x) \ ,$$

dvs funktionen *F* måler arealet under kurven fra $-\infty$ til *x*.

For standardnormalfordelingen har man indført betegnelsen $\Phi(x)$ for den tilhørende fordelingsfunktion:

$$Z \sim N(0,1) \text{ giver, at } \Phi(x) = P(Z \leq x) = \int_{-\infty}^{x} \varphi(t)dt$$

Øvelse 5
Vis som i øvelse 3, at hvis $\Phi_1(x)$ er fordelingsfunktion for normalfordelingen med middelværdi μ og spredning σ, så er $\Phi_1(x) = \int_{-\infty}^{x} \varphi_1(t)dt$, hvor $\varphi_1(t)$ er tæthedsfunktionen hørende til denne normalfordeling

Øvelse 6
I et værktøjsprogram betegner man ofte funktionerne således:

$$\varphi(x) = \text{normpdf}(-\infty, x, 0, 1) \qquad \Phi(x) = \text{normcdf}(-\infty, x, 0, 1)$$

a) Undersøg hvordan dit værktøjsprogram betegner funktionerne.

b) Tegn grafen for $\Phi(x)$.

Opfatter vi grafen for $\Phi(x)$ som en sumkurve, så kan vi spørge, hvilken observation der er knyttet til fx 80%=0,8. Svaret på dette spørgsmål findes ved hjælp af den omvendte funktion til $\Phi(x)$. Den kaldes *den inverse normalfordelingsfunktion* og betegnes ofte invnorm(*x*,0,1).
Fx gælder der: invnorm(0.8,0,1) = 0,841621, fordi $\Phi(0.841621) = 0.8$

c) Opret i et regneark en søjle hørende til 10 %-fraktilerne: 0,1, 0,2, 0,3, …0,9 og lad regnearket beregne en søjle med de tilsvarende x-værdier hørende til standardnormalfordelingen. Opret et punkplot af alle punkterne: (invnorm(0.1,0,1), 0.1), (invnorm(0.2,0,1), 0.2), … (invnorm(0.9,0,1), 0.9) i samme koordinatsystem hvor du har tegnet grafen for $\Phi(x)$.

Det er indlysende, at punkterne i øvelsen ligger på grafen, da vi har konstrueret dem sådan. Men har vi nu i stedet, en række dataværdier, som vi antager er standardnormalfordelt, og opretter vi et punkplot bestående af følgende punkter:

(dataværdiernes højre endepunkter, den kumulerede % i denne værdi)

så skulle punkterne med tilnærmelse ligge på grafen for $\Phi(x)$. Dette kan naturligvis generaliseres til normalfordelinger med middelværdi μ og spredning σ:

Tegner vi sumkurven for et datasæt, som med tilnærmelse er normalfordelt med middelværdi μ og spredning σ, så vil sumkurven med tilnærmelse være lig med fordelingsfunktionen.

Men det betyder omvendt, at vi kan anvende dette til at undersøge en antagelse om normalfordeling. Det kan stadig være svært at se, hvorvidt en given kurve er tæt ved en anden krum kurve. Derfor ønsker vi at transformere situationen, så vi får "strukket" graferne for fordelingsfunktionerne og sumkurverne ud, så de bliver rette linjer! Det gør vi på følgende måde:

Den vandrette x-akse har en normal inddeling. Her afsættes datasættets "højre intervalendepunkter".
Den lodrette y-akse har en dobbelt inddeling, en synlig og en næsten skjult. Den skjulte er en akse med en normal inddeling, og med y=0 midt på aksen
På den synlige akse afsættes 50% i 0. Dernæst afsættes
procenttallet 60 i en afstand fra 0-punktet på $\Phi^{-1}(0.6) = invnorm(0.6,0,1) = 0.253$,
procenttallet 70 i en afstand fra 0-punktet på $\Phi^{-1}(0.7) = invnorm(0.7,0,1) = 0.524$,
procenttallet 80 i en afstand fra 0-punktet på $\Phi^{-1}(0.8) = invnorm(0.8,0,1) = 0.842$,
procenttallet 84,2 i en afstand fra 0-punktet på $\Phi^{-1}(0.842) = invnorm(0.842,0,1) = 1{,}00$
procenttallet 97,7 i en afstand fra 0-punktet på $\Phi^{-1}(0.977) = invnorm(0.977,0,1) = 2{,}00$
procenttallet 99,9 i en afstand fra 0-punktet på $\Phi^{-1}(0.999) = invnorm(0.999,0,1) = 3{,}09$

Procenttal under 50 bliver afsat i tilsvarende afstande i den negative retning af den skjulte akse. Fx afsættes procenttallet 15,8 i en afstand fra 0-punktet på $\Phi^{-1}(0.158) = invnorm(0.158,0,1) = -1{,}00$

Som vi ser, så betyder dette, at enheden på den skjulte akse er spredningen! Læg mærke til at mellem de to spredninger ligger der i alt 84,2%-15,8%= 68,4%, som vi jo ved er gældende for normalfordelinger

Det ligger i selve konstruktionen, at grafen for $\Phi(x)$ er en ret linje med hældning 1: Givet et tal $x = a$ på x-aksen, hvor afsættes den tilhørende værdi på sumkurven, $y = \Phi(a)$? Den afsættes i en afstand fra x-aksen på $\Phi^{-1}(y) = \Phi^{-1}(\Phi(a)) = a$. Dvs ligningen for fordelingsfunktionen bliver $y = x$.

Lad os vende tilbage til vores datasæt. Vi undersøger om dette er normalfordelt ved at se på den såkaldte *standardiserede* z-*score*, der måler afstanden fra middelværdien μ i enheder af spredningen σ. Hvis observationerne *X* stammer fra en stokastisk variabel med middelværdi μ og spredning σ, er den standardiserede z-score derfor givet ved

$$Z = \frac{X - \mu}{\sigma}.$$

Den standardiserede Z-score er derfor centreret omkring 0 og har spredningen 1.
(Læg mærke til, at dette svarer til, at vi ovenfor har isoleret *Z* i ligningen $Y = a \cdot Z + b$, hvor *Z* er standardnormalfordelingen).

Hvis *X* nu rent faktisk er normalfordelt med middelværdi μ og spredning σ, vil *Z* derfor være standardnormalfordelt med middelværdi 0 og spredning 1. Det betyder at ligningen for fordelingsfunktionen for *z* i vores nye koordinatsystem er:

$$y = z = \frac{x - \mu}{\sigma} = \frac{1}{\sigma} \cdot x - \frac{\mu}{\sigma}$$

Dette er en ligning for en linje med hældningskoefficienten $\frac{1}{\sigma}$.
Samtidig ser vi af det første lighedstegn, at hvor $y = 0$ er $x = \mu$.

Fortolkningen af dette er følgende:

- Når vi afsætter punkterne til sumkurven i vores nye kordinatsystem, så vi, de med god tilnærmelse følge en ret linje, hvis *X* er normalfordelt.
- Den rette linje skærer x-asken i $(\mu, 0)$.

- Når vi går σ frem på *x*-aksen går vi 1 op til den rette linje. Men 1 enhed på den skjulte y-akse svarer til den %-afvigelse fra middelværdien, der giver 1 spredning. Derfor kan spredningen aflæses ved at bestemme, hvor linjen skærer den vandrette linje med ligning $y=1$ (i det skjulte koordinatsystem).

Det særlige koordinatsystem, vi har behandlet her, kaldes et *normalfordelingspapir*. Du kan hente et eksmplar af et sådant papir *her.*

Med baggrund i denne fortolkning ser vi, at vi i praksis ikke behøver udregne z-værdierne, men kan gå direkte til de oprindelige dataværdier.

Øvelse 7
Opret nu en tabel med en intervalinddeling af brudstyrkerne og med de kumulerede procenter. Afsæt sumkurven i et normalfordelingspapir og vurder om det er normalfordelt. Aflæs i givet fald middelværdi og spredning, og sammenlign med de tidligere bestemte værdier.

Øvelse 8
En stokastisk variabel *X* er normalfordelt og:
$P(X\geq 7)=0.05$ og $P(X\leq 3)=0.40$
Tegn på et normalfordelingspapir grafen for fordelingsfunktionen for *X*.
Bestem grafisk middelværdi og spredning for *X*.
Bestem $P(X\leq 1.5)$

Øvelse 9
En normalfordelt stokastisk variabel *X* har middelværdi 7 og spredning 2.
Bestem $P(5\leq X\leq 8)$ både ved grafisk metode og ved beregning

Øvelse 10
En virksomhed fremstiller metalplader. Ved kontrol af pladetykkelsen har man fundet, at 1,2% af pladerne har en tykkelse under 14,0mm, og at 95,4% har en tykkelse under 16,0mm.
Det antages at pladetykkelsen er normalfordelt.
Bestem middelværdi og spredning for pladetykkelsen.

Øvelse 11
Ved en sortering af en ærtehøst inddelte man ærterne i tre kategorier: fine, mellemfine og store.
I det følgende antager vi, at diameteren af en ært kan beskrives ved en normalfordelt stokastisk variabel med middelværdi 9,4mm og spredning 1,3mm.

a) Bestem både grafisk og ved beregning hvor mange procent af ærterne, der har en diameter mindre end 7,0mm.

Man foretager inddelingen i de tre kategorier således, at de mindste 30% kaldes fine og de mellemste 45% kaldes mellemfine.

b) Bestem hvilken diameterstørr4else der udgør grænserne mellem de tre typer af ærter.

Øvelse 12

I Øvelse 8.27 i HEM3 er gengivet datamaterialet fra Michelsohn og Newcombes forsøg på at bestemme lysets hastighed. De foretog i juli-september 1882 66 præcisionsmålinger ved hjælp af en forsøgsopstilling, der er beskrevet nærmere i øvelsen.

a) Hent data. Vurder om der er outliers, der skal lægges til side.

I øvelse 8.27 er der gennemført en undersøgelse af dat ved hjælp af histogrammer og sammenligning med en normalfordelingsgraf. Her vil vi inddrage den nye metode:

b) Træk 24800 fra alle tallene og foretag dernæst en gruppering af materialet i intervaller, fx med bredde 1, udregn de kumulerede procenter og tegn en sumkurve i et normalfordelingspapir.

c) Det er rimelig antagelse, at målingerne er normalfordelte. Bestem grafisk middelværdi og spredning.

d) Hent på nettet oplysninger om lysets faktiske hastighed. Hvor stor er afstanden, målt i antal spredninger, mellem Newcombes og Michelsons værdi og den i dag accepterede værdi. Kan forskellen forklares ud fra statistisk usikkerhed, eller skal andre firklaringer inddrages.

I et **projekt 8.9 om *t*-test**, der er et test som anvendes til at vurdere på og sammenligne middelværdier, vender vi tilbage til dette datamateriale, og gennemfører et egentlig test på det som vi vurderer på i punkt d).