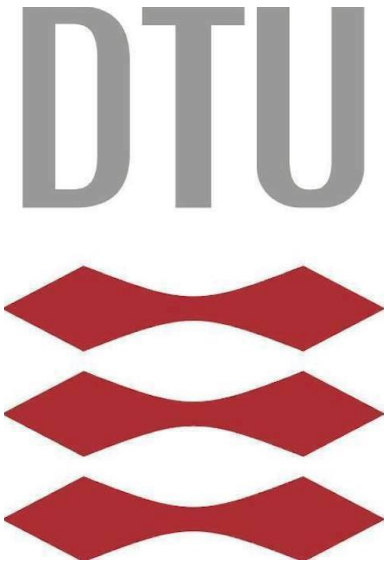
Danmarks Tekniske Universitet



Lukas Leindals
s183920

02403
Introduktion til matematisk statistik

Projekt:
Handel med ETF

18. Juni 2019
Danmarks Tekniske Universitet

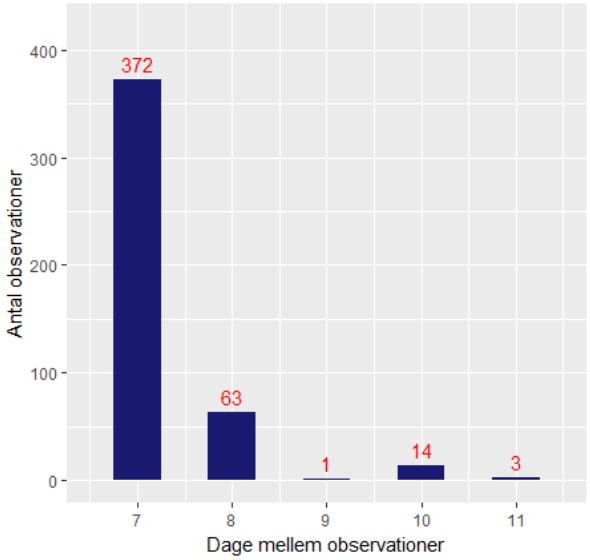
Indhold

Beskrivende analyse

1 Problem 0 - Beskrivende analyse

1.1 Delopgave a

Datasættet *finans1_data* består 95 forskellige ETF'er. For hver ETF er der angivet et relativt ugentligt afkast i perioden 5. Maj 2006 til 8. Maj 2015. Hvis vi kigger på figur 1, viser det sig dog at målinger ikke alle sammen er lavet med 7 dages mellemrum, men har helt op til 11 dage mellem enkelte målinger. Det gælder dog at langt de fleste er taget med 7-8 dages mellemrum, hvilket anses som respektabelt for at vores afkast er ugentligt. Forskellen mellem målinger tyder på at dataen ikke er blevet samlet automatisk og vi kan derfor være ude for at der er manglende data. Ved at undersøge vores data, ses at det ikke er tilfældet. Dette resulterer i at vi ender op med et datasæt bestående af 454 observationer for hver af de 95 ETF'er. Disse obervationer samles i en data frame bestående af en tidskolonne, som angiver hvornår observationen er foretaget, samt 95 kolonner (en for hver ETF), som angiver det relative ugentlige afkast.



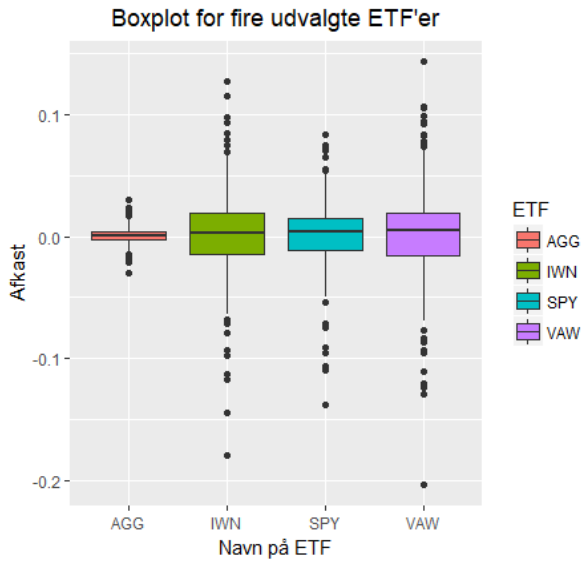
Figur 1: Histogram som viser hvordan antallet af dage mellem hver observation fordeler sig, de røde tal angiver det eksakte antal observationer med den givne tidsforskel mellem observationer

1.2 Delopgave b

EFT	Obs	$\overline{\mu}$	$\overline{\sigma^2}$	$\overline{\sigma}$	Min	Q_1	Median	Q_3	Max
AAG	454	2.658	0.357	59.76	-296.0	-29.73	2.374	38.93	305.1
VAW	454	17.94	13.02	360.8	-2037	-161.0	47.98	196.9	1430
IWN	454	11.88	10.25	320.2	-1797	-143.1	31.20	190.6	1267
SPY	454	13.60	6.143	247.9	-1376	-113.3	42.16	145.0	832.8

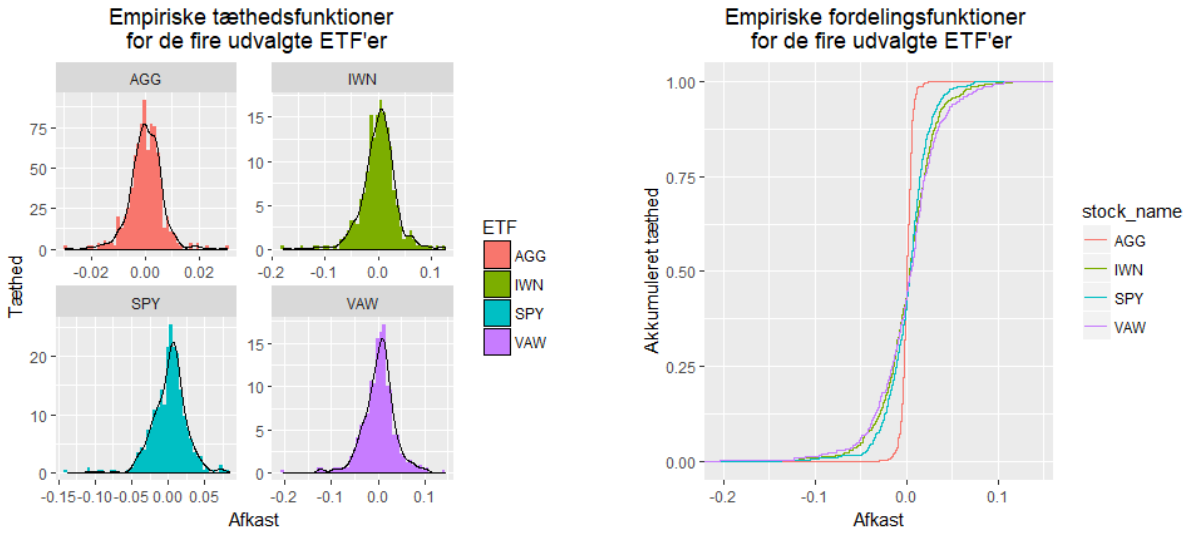
Figur 2: Empiriske værdier for de fire ETF'er *AAG, VAW, IWN* og *SPY*. Dette indebærer antal observationer, middelværdien for stikprøven, variansen for stikprøven, standard afvigelsen for stikprøven, minimumsværdien, nedre kvartil, medianen, øvre kvartil og maximumsværdien. Alle værdier undtagen antal observationer er ganget med 10^4

I tabellen på figur 2 ses en række statistiske værdier for det ugentlige afkast af de fire ETF'er *AAG, VAW, IWN* og *SPY*. Her er det vigtigt at lægge mærke til at værdierne er blevet multipliceret med $10 \cdot 10^4$. En stor del af disse værdier er visuelt repræsenteret i et boxplot på figur 3.



Figur 3: Boxplot for hver af de fire udvalgte ETF'er

På figur 4a ses det empiriske tæthedsfunktioner for de fire ETF'er, både angivet med et histogram, samt en kurve. Summen under denne kurve børe være 1, hvilket også ses på figur 4b, som angiver den fordelingsfunktionerne for hver af de fire ETf'er og dermed deres akkumulerede tæthed.



(a) (b)

Figur 4: Empiriske tætheds- og fordelingsfunktioner for hver af de fire ETF'er

1.3 Delopgave c

På figur 3 og 4b ses det ligesom i tabel 2 at *AGG* adskiller sig en smule fra de tre andre ved at spredningen og variansen er en del mindre og denne ETF må derfor siges at have et stabilt

afkast. Ved at kigge på fordelingsfunktionen, ses denne forskel tydeligt, da de tre andre ETF'er følger omtrent samme mønster, mens *AGG*, nærmest går fra 0 til 1 med det samme. Ud fra tabellen ses at den lave spredning kommer sammen med en lav median og gennemsnit. Dette tyder på at der er en sammenhæng mellem risikoen ved en ETF (variansen/spredningen) og det ugentlige afkast.

Ud fra tabellen på figur 2 ses det at de relative ugentlige afkast for de fire udvalgte ETF'er ligger i intervallet [-0,204; 0,143]. Når vi kigger på figur 3 ses det desuden, at vi har en del outliers i vores datasæt, som bl.a. indebærer vores minimum og maximum. Da vi kigger på det relative ugentlige afkast bør dette ikke have noget at gøre med den svingende tid mellem observationer vi så på figur 1. Når vi kigger på *VAW* ses at denne har to outliers, som ser særligt ekstreme ud, en meget høj og en meget lav. Man kan overveje at fjerne disse fra datasættet, da de kan have en stor påvirkning, men da vi har har 495 observationer, bør det ikke gøre en kæmpe forskel og vi beholder dem, da de trods alt er en del af datasættet og ikke forventes at være en tastefejl. Da både vores minimum og maximum for de fire ETF'er kommer fra *VAW*, må denne siges at ave den største risiko, da den også har den største spredning. Her er der altså meget at vinde, men også meget at tabe.

Når vi kigger på skævheden af de forskellige tæthedsfunktioner, vil det være sådan at for at en venstreskæv fordeling vil have et gennemsnit som er mindre den end medianen, ved at kigge på tabellen fra figur 2, ses at dette gælder for *VAW, IWN* og *SPY*, hvor gennemsnittet er omtrent 3 gange mindre end medianen. Når vi kigger på figur 4a og 4b, ser de dog alle sammen ud til at være nogenlunde symmetrisk fordelt. Vi må derfor konkludere at disse ser ud til at følge en symmetrisk fordeling, men med en lille tendens til at være venstreskæv. For *AGG*, er gennemsnittet en smule større end medianen, når vi kigger på figur 4a og 4b, ser dette dog ikke ud til at resultere i at den er skæv og vi må konkludere at denne er symmetrisk fordelt. Det ser altså ud til at vores data er normal fordelt for de fire ETF'er.

Statistisk analyse 1

2 Problem 1 - Opbygning af porteføljе

Vi går nu fra at kigge på enkelte ETF'er hver for sig, til at kigge på sammensætninger af ETF'er, såkaldte porteføljer.

2.1 Delopgave d

I tabel 1 ses kovariansmatricen for ETF'erne *AGG, VAW, IWN, SPY, EWG* og *EWW*, hvor værdierne er blevet multipliceret med $10 \cdot 10^4$. I diagonalen ses variansen for hver af de 6 ETF'er og det ses at de stemmer overens med værdierne fra tabellen på figur 2, for de ETF'er der går igen. Ud over variansen, ses kovariansen mellem de forskellige ETF'er. En negativ kovaians betyder at det ugentlige afkast for den ene ETF falder, i takt med at det stiger for den anden. Det ses at *AGG* og *IWN* er de to mest uafhængige, mens *VAW* og *EWW* har den største kovarians og *AGG* og *EWG* har den mindste.

	AGG	**VAW**	**IWN**	**SPY**	**EWG**	**EWW**
AGG	0.357	-0.426	-0.259	-0.324	-0.508	-0.371
VAW	-0.426	13.02	9.838	7.927	11.10	11.85
IWN	-0.259	9.838	10.25	7.222	9.502	10.10
SPY	-0.324	7.927	7.222	6.143	8.046	8.153
EWG	-0.508	11.10	9.502	8.046	14.44	11.80
EWW	-0.371	11.85	10.10	8.153	11.80	16.59

Tabel 1: Kovariansmatricen mellem 6 udvalgte ETF'er. Værdierne er ganget med 10^4, og er derfor mindre end vist i tabellen.

2.2 Delopgave e

Vi vælger at kigge på variansen for følgende portføljer bestående af to ETF'er: *(EWG, EWW), (AGG,SPY), (VAW,IWN), (VAW,EWG), (VAW,EWW) og (IWN,EWG).* Disse benævnes *P1, P2, P3, P4, P5* og *P6.* Portføljen laves ved at vælge en andel af den første ETF (α). Portføljerne er derfor givet ved:

$$P = \alpha \cdot ETF1 + (1 - \alpha) \cdot ETF2$$

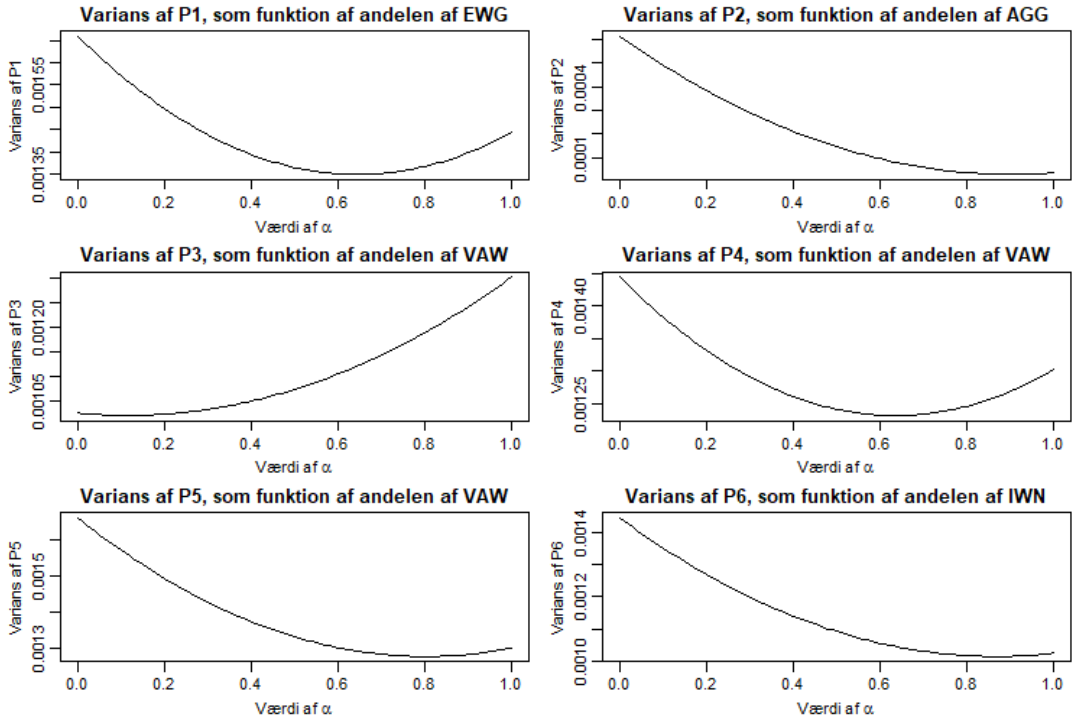
Variansen af sådan en portfølge er givet ved theorem 2.60:

$$\mathrm{Cov}\left(a_0 + a_1X + a_2Y, b_0 + b_1X + b_2Y\right) = a_1b_1\mathrm{V}(X) + a_2b_2\mathrm{V}(Y) + (a_1b_2 + a_2b_1)\,\mathrm{Cov}(X, Y)$$

Da vi er ude efter variansen, tager vi kovariansen med sig selv, således at $a_0 = b_0, a_1 = b_1$ og $a_2 = b2$. Vi får derfor følgende udtryk for variansen:

$$\mathrm{V}(P) = \alpha^2 \cdot \mathrm{V}(ETF1) + (1 - \alpha)^2 \cdot \mathrm{V}(ETF2) + 2\alpha(1 - \alpha) \cdot \mathrm{Cov}(ETF1, ETF2)$$

Variansen for hvert portfølge som funktion af vores andel α, ses på figur 5, hvor vi desuden ligger mærke at hvert plot ser ud til at have et lokalt minimum.



Figur 5: Varians for hvert af de 6 portføljer, som funktion af andelen af den første ETF

Vi ønsker nu at finde den kombination af ETF'er for hvert portfølge, som giver den mindste varians. Altså ønsker vi at optimere vores værdi af α, så vi finder det lokale minimum på figur 5, da vi ser de alle har et lokalt minimum. De 6 portføljer samles i tabel 2, hvor vi desuden ser den optimale værdi af α, samt den minimums varians og den forventede værdi (gennemsnit) denne α værdi resulterer i.

Portfølje	**ETF1**	**ETF2**	**Andel af ETF1** (α)	$\boldsymbol{\mu}$ $[\cdot 10^{-3}]$	**Min**(σ^2) $[\cdot 10^{-4}]$
P1	EWG	EWW	0.645	1.40	13.50
P2	AGG	SPY	0.905	0.37	0.292
P3	VAW	IWN	0.115	1.26	10.20
P4	VAW	EWG	0.635	1.59	12.32
P5	VAW	EWW	0.802	1.78	12.79
P6	IWN	EWG	0.869	1.19	10.15

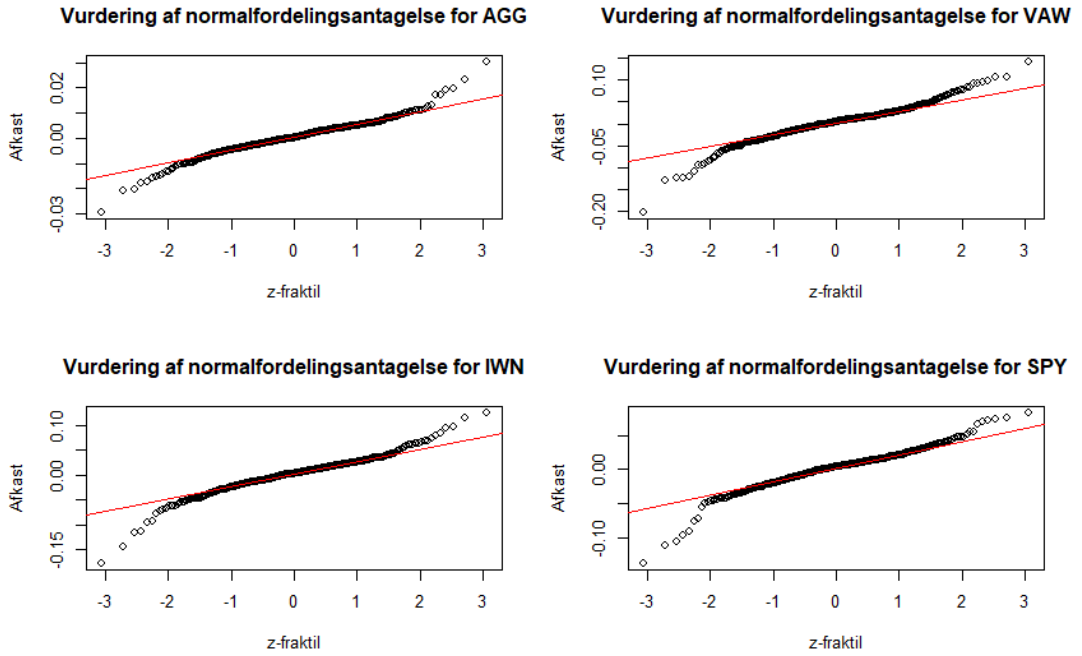
Tabel 2: De 6 portføljer og den værdi af α som giver den mindste varians. Værdierne for variansen og den forventede værdi (gennemsnittet), med denne optimale værdi af α er angivet, efter de er blevet multipliceret med henholdsvis $10 \cdot 10^3$ og $10 \cdot 10^4$

Når vi skal vælge den optimale portføje bør vi gøre os to overvejelser. Første ovevejelse er variansen. Hvis man kigger på denne er *P2* den mest optimale portfølje, da den har mindst varians, den giver til gengæld ikke ik så højt ugentligt afkast. Hvis vi kigger på det ugentlige afkast er *P5* den bedste, da den har højest ugentligt afkast, , den har dog også en af de højeste varianser og er derfor risikofyldt. Vi skal altså afgøre med os selv, hvor stor en risiko vi er villig til at tage, når vi vælger den optimale portfølje. Grundet at *P5* ikke har en varians der er meget større en de fleste af de andre portføljer, vælger vi denne som den optimale portfølje, da den har et højt afkast. Vi vurdere altså at risikoen kun bliver så meget mindre at det påvirker vores beslutning, hvis vi vælger *P2*, men at denne har et for lavt afkast.

3 Problem 2 - Den bedste investering

3.1 Delopgave f

Ud fra tæthedsfordelingerne på figur 4a antager vi at vores data er normal fordelt. Dette testes ved at lave et såkaldt Q-Q plot.



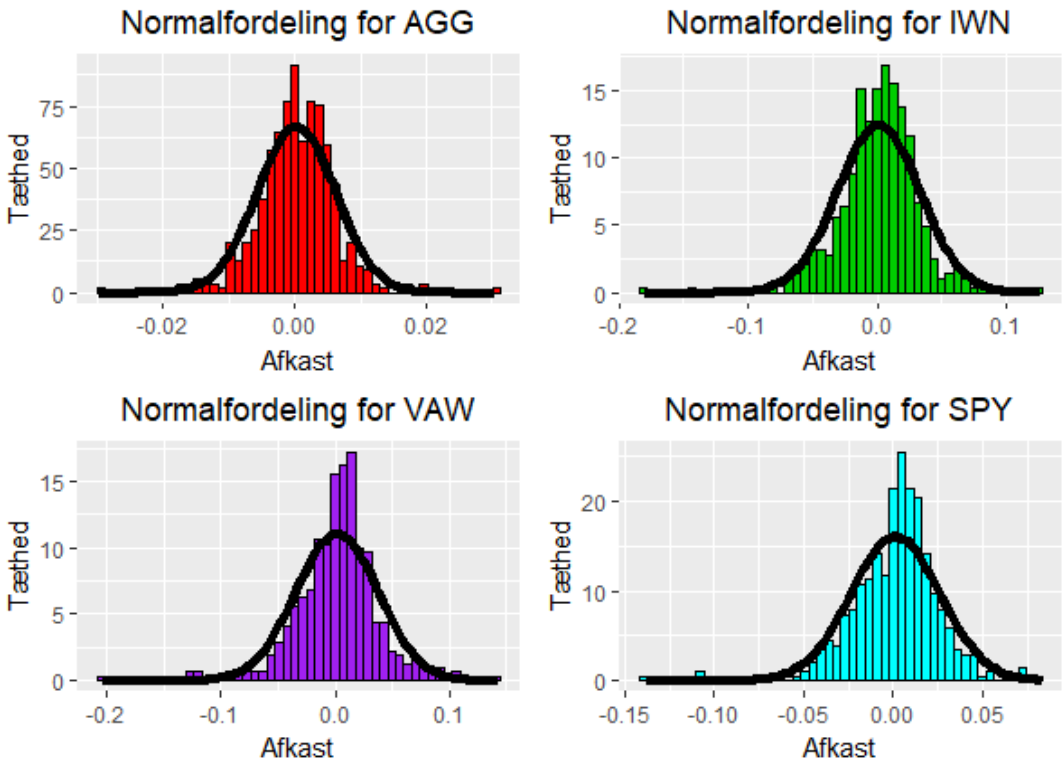
Figur 6: Q-Q plot for de fire ETF'er

På figur 6 ses Q-Q plot for hver af de fire ETF'er. Disse plots viser sammenhængen mellem udvalgte fraktiler og de teoretiske fraktiler, som fremkommer hvis vores data er normalfordelt. Dette betyder at hvis vores data er normal fordelt, vil det følge en ret linje. Denne rette linje er markeret med rød på figur 6. Vi ser at alle de fire ETF'er følger denne rette linje for langt de fleste data punkter og det er kun få outliers, som ikke ligger på linjen. Vi ser derfor at vores antagelse om at vores data er normal fordelt passer godt.

Vi estimere nu modellens parametre. En normal fordeling $\mathcal{N}(\mu, \sigma^2)$ afhænger af gennemsnittet og variansen for vores data. Det gælder for vores data, at hvis det er trukket fra en normalfordeling, vil de mest sandsynlige parameter værdier (most likely) være givet ved det empiriske gennemsnit og den empiriske varians. Vi aflæser disse for hver ETF i tabellen på figur 2 og ser at de ser ud til at følge følgende normalfordelinger:

$$
\begin{aligned}
AGG &\sim \mathcal{N}(2.658 \cdot 10^{-4}, 0.357 \cdot 10^{-4}) \\
VAW &\sim \mathcal{N}(17.94 \cdot 10^{-4}, 13.02 \cdot 10^{-4}) \\
IWN &\sim \mathcal{N}(11.88 \cdot 10^{-4}, 10.25 \cdot 10^{-4}) \\
SPY &\sim \mathcal{N}(13.60 \cdot 10^{-4}, 6.143 \cdot 10^{-4})
\end{aligned}
$$

På figur 7 har vi plottet disse normalfordelinger sammen med vores tæthedsfunktioner. Vi ser at disse normalfordelinger følger vores empiriske tæthedsfunktioner til en hvis grad. Denne forskel kan bl.a. skyldes antallet af bins, som angiver hvor store intervallerne for hver søjle er. Vi ser altså også her en lille forskel, som vi så det med vores Q-Q plot. Trods dette ser vores data, stadig ud til at være forholdsvis normalfordelt.



Figur 7: Tæthedsfunktioner for de fire ETf'er sammen med deres tilhørende nomalfordelinger vist med sort kurve.

3.2 Delopgave g

95 % konfidensintervallet for det gennemsnitlige ugentlige afkast er givet ved, under antagelsen af at det er normal fordelt:

$$\overline{x} \pm t_{0.975} \cdot \frac{s}{\sqrt{n}}$$

hvor $\overline{x}$ er det gennemsnilige ugentlige afkast, $\pm t_{0.975}$ er den 97,5'ne fraktil for en t-fordeling, s er standardafvigelsen og n er antal observationer. Vi bestemmer t-fordelingsfraktilen til 1.985 og resten af værdierne aflæses i tabel 2:

$$\mu_{AGG} = 2.658 \cdot 10^{-4} \pm 1.985 \cdot \frac{59.76 \cdot 10^{-4}}{\sqrt{454}} \longrightarrow \mu_{AGG} \in [-2.854, 8.169] \cdot 10^{-4}$$

$$\mu_{VAW} = 17.94 \cdot 10^{-4} \pm 1.985 \cdot \frac{360.8 \cdot 10^{-4}}{\sqrt{454}} \longrightarrow \mu_{VAW} \in [-15.34, 51.22] \cdot 10^{-4}$$

$$\mu_{IWN} = 11.88 \cdot 10^{-4} \pm 1.985 \cdot \frac{320.2 \cdot 10^{-4}}{\sqrt{454}} \longrightarrow \mu_{IWN} \in [-17.65, 41.41] \cdot 10^{-4}$$

$$\mu_{SPY} = 13.60 \cdot 10^{-4} \pm 1.985 \cdot \frac{247.9 \cdot 10^{-4}}{\sqrt{454}} \longrightarrow \mu_{SPY} \in [-9.260, 36.46] \cdot 10^{-4}$$

Disse værdi stemmer som forventet overens med de værdier som opnås i R med *t.test*. Vores middelværdi med et konfidensinterval på 95 %, ses altså kan svinge en del, vi kan faktisk risikere at vores gennemsnit er negativt. Vi kan altså ikke med 95 % sikkerhed sige at vores gennemsnit er positivt. Vi skal derfor bruge mere data, hvis vi skal være mere sikre på hvilke ETF'er, som giver et positivt afkast.

Da alle ETF'erne har samme antal observationer, er det kun standard afvigelsen, som bestemmer bredden af vores interval, vi bestemmer derfor et konfidensinterval for vores varians

på 95 %. Dette interval bestemmes ved:

$$\sigma^2 \in \left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}; \frac{(n-1)s^2}{\chi^2_{\alpha/2}} \right]$$

her er forskellen fra før at vi bruger en χ^2-fordeling, hvor α er givet ved 0.05, når vi har et konfidens interval på 95 %. Vi beregner for hver af vores ETF'er:

$$\sigma^2_{AGG} \in \left[\frac{(454-1)\cdot 0.357\cdot 10^{-4}}{123.9}; \frac{(454-1)\cdot 0.357\cdot 10^{-4}}{69.92} \right] \longrightarrow \sigma^2_{AGG} \in [0.315, 0.409]\cdot 10^{-4}$$
$$\sigma^2_{VAW} \in \left[\frac{(454-1)\cdot 13.02\cdot 10^{-4}}{123.9}; \frac{(454-1)\cdot 13.02\cdot 10^{-4}}{69.92} \right] \longrightarrow \sigma^2_{VAW} \in [11.48, 14.90]\cdot 10^{-4}$$
$$\sigma^2_{IWN} \in \left[\frac{(454-1)\cdot 10.25\cdot 10^{-4}}{123.9}; \frac{(454-1)\cdot 10.25\cdot 10^{-4}}{69.92} \right] \longrightarrow \sigma^2_{IWN} \in [9.036, 11.73]\cdot 10^{-4}$$
$$\sigma^2_{SPY} \in \left[\frac{(454-1)\cdot 6.143\cdot 10^{-4}}{123.9}; \frac{(454-1)\cdot 6.143\cdot 10^{-4}}{69.92} \right] \longrightarrow \sigma^2_{SPY} \in [5.416, 7.029]\cdot 10^{-4}$$

For varianserne ser vi ligesom med middelværdierne, at intervallerne ikke er lige store, da vi her skalerer med varianserne. Vi ser desuden at der ikke er nær så stor bredde på intervallet, som for middelværdierne og det snævre interval betyder at vi er ret sikre på den varians vi ser i tabellen på figur 2.

3.3 Delopgave h

Vi vil nu forsøge at bestemme vores konfidensintervaller vha. non-parametric bootstrap. Her trækker vi 454 værdier for hver ETF, ud fra de værdier af afkast, som vi har for hver uge (med tilbagelægning). Vi sætter vores seed til 7285, så resultaterne er reproducerbare. Dette gøres 10000 gange, hvorefter vi beregner det forventede ugentlige afkast og varians, så vi i alt har 10000 middelværdier og varianser. Vi kan nu bestemme et 95 % konfidensinterval ved at bestemme den 2.5 og 97.5 fraktil af vores middelværdier og varianser. Resultaterne ses i tabel 3 sammen med de resultater vi opnåede i delopgave g. Vi ser at middelværdierne minder meget om hinanden og at varianserne også gør det, men at varianserne variere en lille smule fra vores t-fordelte resultater. Da vores bootstrap resultater, som ikke har nogen forudantagede parametre, ligger så tæt på vores t-fordelte resultater, må vi kunne konkludere at det var en god antagelse at vores data er normal fordelt.

	Konfidens intervaller for $\mu \cdot 10^{-4}$		**Konfidensintervaller for $\sigma^2 \cdot 10^{-4}$**	
	t-fordelt	**Bootstrap**	**t-fordelt**	**Bootstrap**
AGG	[-2.854, 8.169]	[-2.863, 8.112]	[0.315, 0.409]	[0.284, 0.441]
VAW	[-15.34, 51.22]	[-15.52, 50.75]	[11.48, 14.90]	[10.36, 16.10]
IWN	[-17.65, 41.41]	[-17.85, 40.50]	[9.036, 11.73]	[8.085, 12.75]
SPY	[-9.260, 36.46]	[-9.278, 36.01]	[5.416, 7.029]	[4.818, 7.629]

Tabel 3: Sammenligning af konfidensintervaller for det forventede ugentlige afkast og variansen, når vi anvender henholdsvis værdien beregnet vha. en t-fordeling samt non-parametric bootstrap.

3.4 Delopgave i

Vi tester nulhypotesen:

$$H_0 : \mu_{ETF} = 0$$

Dette gøres ved at beregne p-værdien. Denne kan findes ved:

$$p \text{ -value } = 2 \cdot P\left(T > |t_{\text{ obs }}|\right)$$

hvor T følger en t-fordeling med n-1 frihedsgrader og t_{obs} beregnes ved:

$$t_{\text{obs}} = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

Hvor μ_0 er 0 ifølge vores nulhypotese. Vi får følgende p-vædier:

ETF	**p-værdi**
AGG	0.344
VAW	0.290
IWN	0.430
SPY	0.243

Tabel 4: Tabel over p-værdierne for de fire ETF'er

Det ses fra tabel 4 at alle p-værdierne ligger et godt stykke over 0.1 og der er derfor ikke noget signifikant bevis mod vores nulhypotese og vi kan derfor ikke afkræfte den. Det ser derfor ud som om at vi ikke opnår det store ved at investere i en ETF kontra at gemme vores penge under hovedpuden, når vi kigger på vores hypotese test

4 Problem 3

4.1 Delopgave j

Vi undersøger nu om der er et homogent afkast mellem de 4 ETF'er, ved at undersøge følgende nulhypotese:

$$H_0 : \mu_{ETF_{low}} = \mu_{ETF_{high}}$$

De to ETF'er med henholdsvis lavest og højest gennemsnitligt afkast, aflæses fra tabellen på figur 2 til at være *AGG* og *VAW*. Vi bestemmer nu p-værdien for en two-sample t-test i *R*, med kommandoen *t.test* og et konfidens interval på 95 %. Dette giver en p-værdi på 0.374, hvilket er godt over 0.1 og vi har derfor ikke noget signifikant bevis mod vores nulhypotese, hvilket betyder vi ikke kan afkræfte den samt at det tyder på der er et homogent afkast mellem ETF'erne

Statistisk analyse 2

5 Problem 4

Vi kigger nu på datasættet *finans2*, som indeholder følgende 7 variable for hver af de 95 ETF'er:

Variabel	Betydning	Enhed
X	Navn på ETF	
Geo.mean	Geometrisk gennemsnits relative ugentlige afkast r_{uge}	Pct.
Volatility	Den ugentlige volatilitet	Pct.
maxDD	Maximum Draw Down	Pct.
maxTuW	Maximum Time under Water	Pct.
VaR	Ugentligt Value-at-Risk	Pct.
CVaR	Ugentligt Conditional Value at Risk	Pct.

Figur 8: De tilgængelige variable fra *finans2*

5.1 Delopgave k

Vi kigger på korrelationen mellem de forskellige risikomål. Korrelationen mellem to variable kan bestemmes ved følgende formel:

$$\text{Cor}(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$$

Eksempelvis beregnes den for Geo.mean og maxTuW til følgende:

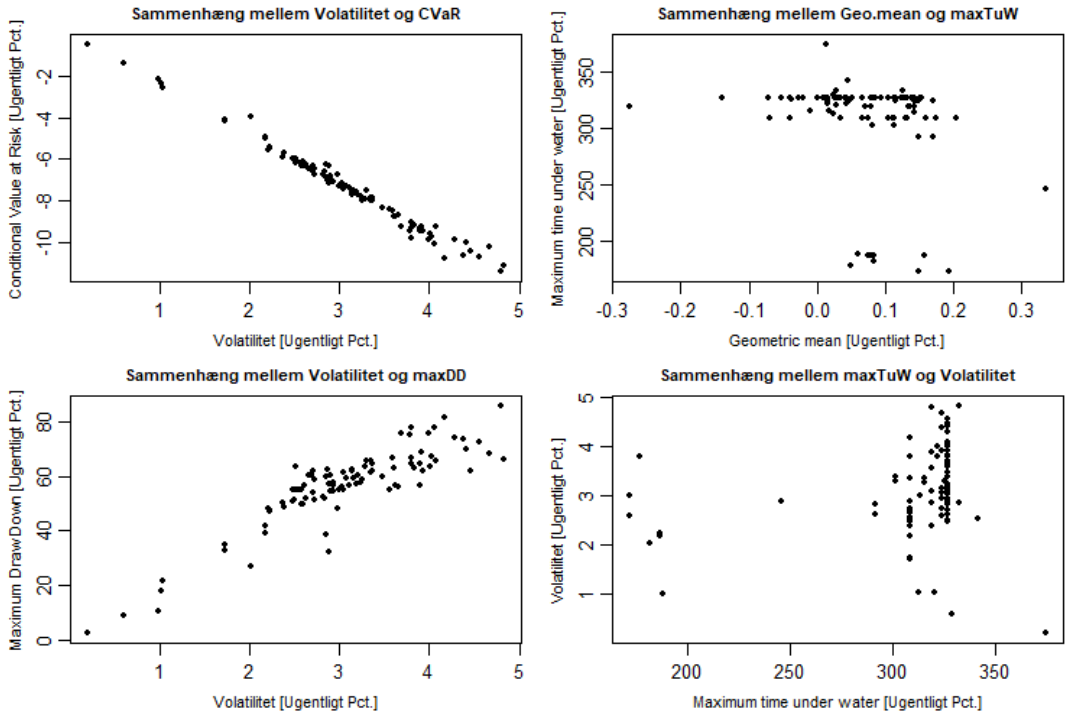
$$\text{Cor}(Geo.mean, maxTuW) = \frac{\text{Cov}(Geo.mean, maxTuW)}{\sigma_{Geo.mean} \cdot \sigma_{maxTuW}} = \frac{-0.753}{80.9 \cdot 10^{-3} \cdot 42.8} = -0.218$$

Dette gøres for samtlige kombinationer og resultaterne samles i tabel 5 og som forventet får vi samme værdi, som den vi beregner med R. Vi bemærker at vi i diagonalen har 1 hele vejen, hvilket er forventeligt, da dette svarer til korrelationen med sig selv.

	Geo.mean	**Volatilitet**	**maxDD**	**maxTuW**	**VaR**	**CVaR**
Geo.mean	1	-0.357	-0.397	-0.218	0.411	0.384
Volatilitet	-0.357	1	0.880	0.251	-0.973	-0.992
maxDD	-0.397	0.880	1	0.281	-0.863	-0.909
maxTuW	-0.218	0.251	0.281	1	-0.267	-0.264
VaR	0.411	-0.973	-0.863	-0.267	1	0.968
CVaR	0.384	-0.992	-0.909	-0.264	0.968	1

Tabel 5: Korrelationen mellem de 6 numeriske variable fra *finans2*

Vi udvælger nu 4 kombinationer af risiko variable, som vi kigger nærmere på volatilitet og CVaR & volatilitet og maxDD, som begge har en stærk korrelation samt Geo.mean og maxTuW & maxTuW og volatilitet, som begge ser ud til nærmest ikke at være korrelerede ifølge tabel 5. Vi plotter de fire kombinationer, for at få et visuelt billed af sammenhængen:



Figur 9: Sammenhængen mellem 4 par af risiko varaible, 2 med høj korrelation og 2 med lav korrelation

På figur 9, ses det tydeligt hvilket to plots, som er dem der har en høj korrelation, da vi ser en kraftig lineær tendens for dem begge. Denne tendens er opadgående for den positive korrelation og nedadgående for den negative korrelation. Samtidig ser vi at vi ikke rigtig kan sige noget den ene variables værdi ud fra den anden, når vi kigger på plotsne, for de varaible som ikke har en høj korrelation.

5.2 Delopgave l

Vi opstiller en simpel lineær regressionsmodel med responsvariablen Geo.mean og VaR, da denne har den højeste korrelation med Geo.mean (jf tabel 5) og dermed må forventes at give den bedste lineære model:

$$Geo.mean_i = \beta_0 + \beta_1 VaR_i + \varepsilon_i$$

hvor β_0 er værdien af Geo.mean, når VaR er 0, β_1 er den den procentdel det geometriske gennemsnit vokser med, når vores value at risk stiger 1 % og den stokastiske variabel ϵ_i er afvigelsen fra den empiriske værdi af Geo.mean. Denne afvigelse formodes at følge en normalfordeling:

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

hvor σ^2 kaldes residual standard error og beskriver hvor meget $epsilon_i$ afviger.

5.3 Delopgave m

Vi ønsker nu at bestemme koefficienterne β_0 og β_1 i vores model. Dette gøres vha. least squares estimaterne, som minimere summen af de kvadrerede afvigelser:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{S_{xx}}$$
$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

hvor $S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$

Vi beregner nu de forskellgie værdier og indsætter dem:

$$\hat{\beta}_1 = \frac{4.415}{187.6} = 23.53 \cdot 10^{-3}$$
$$\hat{\beta}_0 = 76.9 \cdot 10^{-3} 23.53 \cdot 10^{-3} \cdot -4.655 = 0.186$$

Vi får altså vores lineære model til at være følgende:

$$Geo.mean_i = 0.186 + 23.53 \cdot 10^{-3} VaR_i$$

Det vil altså sige at vores geometriske gennemsnits relative ugentlige afkast er 0.186 % når vores value at risk er 0 % og at dette gennemsnit vokser med $23.53 \cdot 10^{-3}$%, når value at risk vokser med 1 %. Vi ønsker nu at vurdere, hvor stor den fejl vi begår ved vores model er. Dette gøres ved at bestemme modelvariansen $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{RSS\left(\hat{\beta}_0, \hat{\beta}_1\right)}{n-2}$$

hvor n er antallet af observationer og RSS (residual sum of squares) er den kvadrerede afstand mellem vores model og vores empiriske værdier:

$$RSS\left(\beta_0, \beta_1\right) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(Geo.mean_i - (\beta_0 + \beta_1 VaR_i)\right)^2$$

Vi beregner nu modelvariansen:

$$\hat{\sigma}^2 = \frac{0.511}{95-2} = 5.49 \cdot 10^{-3}$$

Vores endelige lineære model bliver altså:

$$Geo.mean_i = 0.186 + 23.53 \cdot 10^{-3} VaR_i + \epsilon_i$$

hvor fejlen ϵ_i følger normalfordelingen $\epsilon_i \sim \mathcal{N}(0, 5.49 \cdot 10^{-3})$

5.4 Delopgave n

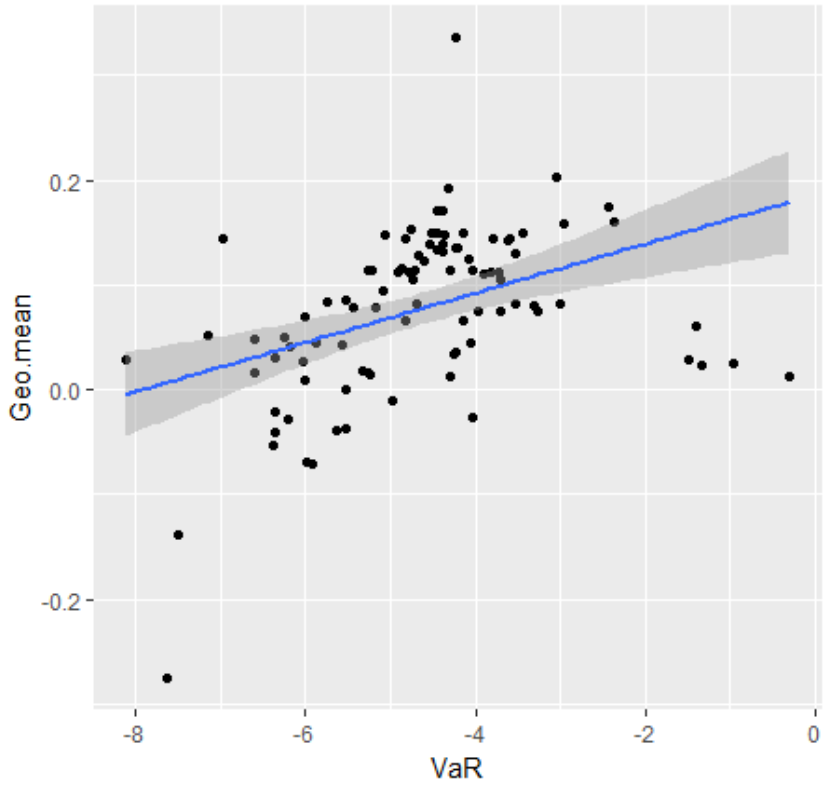
Men hvor meget af variansen er forklaret med modellen? Dette kan vi bestemme ved at kigge på den forklarede varians, som er den andel af variansen der er forklaret af modellen. Denne beregnes således:

$$r^2 = 1 - \frac{\sum_i \left(Geo.mean_i - \hat{Geo.mean_i}\right)^2}{\sum_i \left(Geo.mean_i - \overline{Geo.mean}\right)^2}$$

hvor $\hat{Geo.mean_i} = \hat{\beta}_0 + \hat{\beta}_1 VaR_i$

Denne aflæses i R til 0.169 og dette giver en forklaret korrelation på $\sqrt{0.169} = 0.411$ og der er derfor ikke store tegn på at der findes en korrelation.

En anden måde at vurdere om der er korrelation er ved at kigge på en hypotese test hvor vores nulhypotese er, at hældningen er 0. Hvis dette kan afkræftes siges der er være en korrelation mellem de to variable. Vi aflæser p-værdien for denne hypotese til $3.50e \cdot 10^{-5}$, hvilket er langt under det sædvanlige signifikansniveau på 5 %. Det tyder altså her på at der er en korrelation mellem de to variable. Vi har altså to resultater, som peger lidt i hver sin retning. Vi plotter derfor vores data for at foretage den endelige konklusion:



Figur 10: De empiriske datapunkter for Geo.mean og VaR sammen med en linje der repræsentere vores model og standardafvigelserne for denne model.

Når vi kigger på figur 10 ser at de færreste punkter passer med vores model. Selv når vi kigger på standardafvigelserne for modellen (det grå område) indeholder vores model ikke så mange af vores datapunkter og vi må derfor konkludere, at der ikke er nogen signifikant korrelation mellem Geo.mean og VaR.