TMA4267 - Linear Statistical Models
Exercise 2 - V2014

Problem 1: The chi-square, t and F-distribution

a) $U \sim N(0,1)$. Find pdf and MGF of $X = U^2$.

Denote by ϕ the pdf of the standard Normal distribution.
Let $X = U^2$ and $U = \sqrt{X}$.

$$\begin{aligned}
F_X(x) &= P(U^2 \leq x) = P(-\sqrt{x} \leq U \leq \sqrt{x}) = F_U(\sqrt{x}) - F_U(-\sqrt{x}) \\
f_X &= \frac{d}{dx}F_X(x) = f_U(\sqrt{x})\frac{d}{dx}\sqrt{x} - f_U(-\sqrt{x})\frac{d}{dx}(-\sqrt{x}) \\
&= f_U(\sqrt{x})\frac{1}{2\sqrt{x}} + f_U(-\sqrt{x})\frac{1}{2\sqrt{x}} \\
&= \frac{1}{\sqrt{2\pi}}e^{-x/2}\frac{1}{2\sqrt{x}} + \frac{1}{\sqrt{2\pi}}e^{-x/2}\frac{1}{2\sqrt{x}} \\
&= \frac{1}{\sqrt{2\pi}}e^{-x/2}x^{-1/2} \\
&= \frac{1}{\sqrt{2}\Gamma(1/2)}e^{-x/2}x^{1/2-1}
\end{aligned}$$

MGF:

$$\begin{aligned}
M_{U^2}(t) &= \int_{-\infty}^{\infty} e^{tu^2}\phi(u)du = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tu^2}e^{-u^2/2}du \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-u^2(1-2t)/2}du \text{ using } u = v(1-2t), du = (1-2t)dv \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-v^2/2}\frac{1}{\sqrt{1-2t}}dv \\
&= \frac{1}{\sqrt{1-2t}} \text{ for } t < \frac{1}{2}
\end{aligned}$$

b) $V \sim \chi^2_p$.

First we use the result from a) to find the MGF for the χ^2_p. Since V can be formed by a sum of p independent χ^2_1 variables, then the MGF of V is the product of the MGF of p χ^2_1 variables.

$$M_V(t) = [M_U^2(t)]^p = \frac{1}{(1-2t)^{p/2}}$$

Then we find the MGF of V directly from $f_V(v)$.

$$\begin{aligned}
M_V(t) &= \int_{-\infty}^{\infty} e^{tv} \frac{1}{\Gamma(p/2)2^{p/2}} v^{(p/2)-1} e^{-v/2} dv \\
&= \frac{1}{\Gamma(p/2)2^{p/2}} \int_{-\infty}^{\infty} e^{-v/2(1-2t)} v^{p/2-1} dv, \text{ let } u = v(1-2t), du = (1-2t)dv \\
&= \frac{1}{\Gamma(p/2)2^{p/2}} \int_{-\infty}^{\infty} e^{-u/2} \frac{u^{p/2-1}}{(1-2t)^{p/2-1}} \frac{du}{(1-2t)} \\
&= \frac{1}{(1-2t)^{p/2}} \frac{1}{\Gamma(p/2)2^{p/2}} \int_{-\infty}^{\infty} e^{-u/2} u^{p/2-1} du \\
&= \frac{1}{(1-2t)^{p/2}}
\end{aligned}$$

(The last integral equals 1 since the integrand is the χ^2-distribution.) We see that the two calculations of $M_V(t)$ are equal, and thus conclude that the given $f_V(v)$ is for the χ^2_p-distribution.

c) Let $V \sim \chi^2_p$ and $W \sim \chi^2_q$, where V and W are independent. The joint distribution is then the product of the two marginal distributions.

$$f_{V,W}(u,v) = \frac{1}{\Gamma(p/2)2^{p/2}} v^{(p/2)-1} e^{-v/2} \cdot \frac{1}{\Gamma(q/2)2^{q/2}} w^{(q/2)-1} e^{-w/2}$$

Let then $F = \frac{V/p}{W/q}$, and $G = W$, and use the multivariate transformation formula to find the joint pdf of F and G. We start with the inverse functions and the Jacobian.

$$\begin{aligned}
V &= \frac{p}{q} FG \\
W &= G \\
J &= \det \begin{bmatrix} \frac{p}{q} g & \frac{p}{g} f \\ 0 & 1 \end{bmatrix} = \frac{p}{q} g \\
f_{F,G}(f,g) &= f_{V,W}(\frac{p}{q} fg, g) \cdot \frac{p}{q} g \\
&= \frac{1}{\Gamma(p/2)\Gamma(q/2)2^{(p+q)/2}} (\frac{p}{q} fg)^{(p/2)-1} g^{q/2-1} e^{-(\frac{p}{q}f+1)g/2} \frac{p}{q} g \\
&= \frac{1}{\Gamma(p/2)\Gamma(q/2)2^{(p+q)/2}} (\frac{p}{q})^{p/2} f^{p/2-1} g^{(p+q)/2-1} e^{-(\frac{p}{q}f+1)g/2}
\end{aligned}$$

Then, find the marginal distribution of F from this joint distibution.

$$\begin{aligned}
f_F(f) &= \frac{1}{\Gamma(p/2)\Gamma(q/2)2^{(p+q)/2}}(\frac{p}{q})^{p/2}f^{p/2-1}\int_0^\infty g^{(p+q)/2-1}e^{-(\frac{p}{q}f+1)g/2}dg \\
u &= (\frac{p}{q}f+1)g \text{ and } du = (\frac{p}{q}f+1)dg \\
f_F(f) &= \frac{1}{\Gamma(p/2)\Gamma(q/2)2^{(p+q)/2}}(\frac{p}{q})^{p/2}f^{p/2-1}\int_0^\infty \frac{u^{(p+q)/2-1}}{(\frac{p}{q}f+1)^{(p+q)/2-1}}e^{-u/2}\frac{du}{\frac{p}{q}f+1} \\
&= \frac{(\frac{p}{q})^{p/2}f^{p/2-1}}{\Gamma(p/2)\Gamma(q/2)2^{(p+q)/2}(\frac{p}{q}f+1)^{(p+q)/2}}\int_0^\infty u^{(p+q)/2-1}e^{-u/2}du \\
&= \frac{2^{(p+q)/2}\Gamma(\frac{p+q}{2})(\frac{p}{q})^{p/2}f^{p/2-1}}{2^{(p+q)/2}\Gamma(p/2)\Gamma(q/2)}(\frac{p}{q}f+1)^{(p+q)/2}\int_0^\infty \frac{1}{2^{(p+q)/2}\Gamma(\frac{p+q}{2})}u^{(p+q)/2-1}e^{-u/2}du \\
&= \frac{\Gamma(\frac{p+q}{2})(\frac{p}{q})^{p/2}}{\Gamma(p/2)\Gamma(q/2)}\frac{f^{p/2-1}}{(\frac{p}{q}f+1)^{(p+q)/2}}
\end{aligned}$$

d) Let $U \sim N(0,1)$ and $V \sim \chi^2_p$, and U and V are independent. Find the pdf of the random variable $T = \frac{U}{\sqrt{V/p}}$.

First, the joint pdf of U and V, by multiplying the marginal pdfs.

$$f_{U,V}(u,v) = \frac{1}{\sqrt{2\pi}}e^{-u^2/2}\cdot\frac{1}{\Gamma(p/2)2^{p/2}}v^{(p/2)-1}e^{-v/2}$$

Now , the inverse of the transformation $t = \frac{u}{\sqrt{v/p}}$ and $w = v$ is $u = t\sqrt{w/p}$ and $v = w$, with Jacobian $\sqrt{w/p}$. This gives joint distribution $f_{T,W}(t,w)$:

$$f_{T,W}(t,w) = f_{U,V}(t(\sqrt{w/p}),w)\cdot\sqrt{w/p}$$

The marginal pdf of T is

$$\begin{aligned}
f_T(t) &= \int_0^\infty f_{U,V}(t\sqrt{w/p},w)\cdot\sqrt{w/p}\,dw \\
&= \frac{1}{(2\pi)^{1/2}\Gamma(\frac{p}{2})2^{p/2}}\int_0^\infty e^{-(1/2)t^2w/p}w^{p/2-1}e^{-w/2}\left(\frac{w}{p}\right)^{1/2}dw \\
&= \frac{1}{(2\pi)^{1/2}\Gamma(\frac{p}{2})2^{p/2}p^{1/2}}\int_0^\infty e^{-(1/2)(1+t^2/p)w}w^{(p+1)/2-1}dw
\end{aligned}$$

The trick now it to recognice that the integrand is the pdf of the gamma distribution $\frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-x/\beta}$, with parameters $\alpha = (p+1)/2$ and $\beta = 2/(1+t^2/p)$, so the integral is 1.

$$\begin{aligned}
f_T(t) &= \frac{\Gamma(\frac{p+1}{2})(\frac{2}{1+t^2/p})^{(p+1)/2}}{(2\pi)^{1/2}\Gamma(\frac{p}{2})2^{p/2}p^{1/2}}\int_0^\infty \frac{1}{\Gamma(\frac{p+1}{2})(\frac{2}{1+t^2/p})^{(p+1)/2}}e^{-1/(1+t^2/p)}\,w^{(p+1)/2-1}dw \\
&= \frac{\Gamma(\frac{p+1}{2})}{\Gamma(\frac{p}{2})}\frac{1}{(p\pi)^{1/2}}\frac{1}{(1+t^2/p)^{(p+1)/2}}
\end{aligned}$$

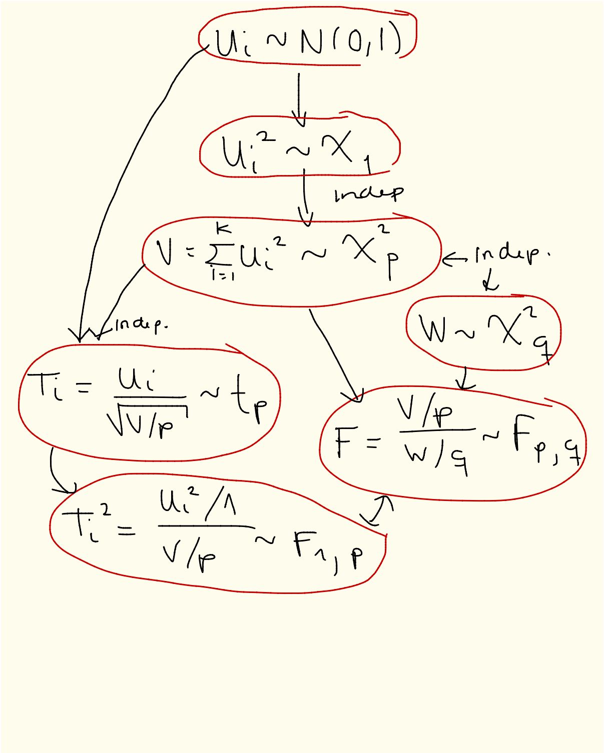
which is the T pdf.

e) Let $T \sim t_q$ (t-distribution with q degrees of freedom). Then show that $T^2 \sim F_{1,q}$.

$$
\begin{aligned}
T^2 &= \left(\frac{U}{\sqrt{V/p}}\right)^2 \\
&= \frac{U^2/1}{V/p}
\end{aligned}
$$

We see that the numerator is χ_1^2 and the denominator is χ_p^2, and from 1d we see that T^2 is Fisher with 1 and p degrees of freedom.

f)


$U_i \sim N(0,1)$
$U_i^2 \sim \chi_1$
indep
$V = \sum_{i=1}^{k} U_i^2 \sim \chi^2_p$
indep.
$W \sim \chi^2_q$
indep.
$T_i = \frac{U_i}{\sqrt{V/p}} \sim t_p$
$F = \frac{V/p}{W/q} \sim F_{p,q}$
$T_i^2 = \frac{U_i^2/1}{V/p} \sim F_{1,p}$

Problem 2: N, Chi-square, t and F by simulation - in R

```
# Problem 2, only R.

B <- 10000
n <- 10

# a
rnorm(B,0,1) # draw B standard normal variates
dchisq(1,1) # density at x=1 for chi-square df=1
pt(0,n-1) # cdf at x=0 for t-distr with df=n-1
qf(0.05,1,2) #critical value with area 0.05 to the left
qf(0.05,1,2,lower.tail=FALSE) # critical value with area 0.05 to the right
qf(0.95,1,2) # same as above

# b
?curve
# how far out? 4 sds ok?
curve(dnorm,-4,4,type="l")
abline(v=qnorm(0.05),col=2)
abline(v=qnorm(0.95),col=2)
# for the fun of it, adding shades to tails
tt <- seq(from = -4, to=qnorm(0.05), length = 50)
dtt <- dnorm(tt)
polygon(x = c(-4, tt, qnorm(0.05)), y = c(0, dtt, 0), col = "gray")
tt <- seq(from = qnorm(0.95), to=4, length = 50)
dtt <- dnorm(tt)
polygon(x = c(qnorm(0.95),tt,4), y = c(0, dtt, 0), col = "gray")

# c
x <- rnorm(B,0,1)
y <- x^2
range(y)
hist(y,nclass=100,prob=TRUE)
dchisq1 <- function(x) return(dchisq(x,df=1))
curve(dchisq1,min(y),max(y),add=TRUE,col=2)
# curve only takes a function with ONE argument, needed to make a df=1 version of dchisq
abline(v=qchisq(0.1,1),col=3)
abline(v=qchisq(0.9,1),col=3)

# d
x <- rnorm(B)
y <- rchisq(B,df=n-1)
t <- x/sqrt(y/(n-1))
hist(t,nclass=50,prob=TRUE)
dt9 <- function(x) return(dt(x,df=9))
curve(dt9,min(t),max(t),add=TRUE,col=2)
# alternative to curve - plot two vectors
xvec <- seq(min(t),max(t),length=100)
yvec <- dt(xvec,df=n-1)
lines(xvec,yvec,col=4)
abline(v=qt(0.15,n-1),col=5)
abline(v=qt(0.85,n-1),col=5)

# e
f <- t^2
hist(f,nclass=50,prob=TRUE)
xvec <- seq(min(f),max(f),length=100)
yvec <- df(xvec,1,n-1)
lines(xvec,yvec,col=2)
abline(v=qf(0.05,1,n-1),col=5)
abline(v=qf(0.95,1,n-1),col=5)

# f more F
n1 <- 5
n2 <- 40
u <- rchisq(B,df=n1)
v <- rchisq(B,df=n2)
f <- u*n2/(v*n1)
hist(f,nclass=50,prob=TRUE)
xvec <- seq(min(f),max(f),length=100)
yvec <- df(xvec,n1,n2)
lines(xvec,yvec,col=2)
abline(v=qf(0.05,n1,n2),col=3)
abline(v=qf(0.95,n1,n2),col=3)
```

Problem 3: Teorem 2.4 of Bingham & Fry

```
# Problem 3: numerical study of normal samples

B <- 10000
n <- 20
# choosing som starting parameters
mu <- 1
sigma <- 2
xmat <- matrix(rnorm(B*n,mu,sigma),ncol=n)
mean(c(xmat))
sd(c(xmat))

xbar <- apply(xmat,1,mean) # vector of B means
s2 <- apply(xmat,1,var)*(n-1)/n # default n-1 in var

#i independence of xbar and s2
plot(xbar,s2) #looks like random scatter
summary(lm(s2~xbar)) # fit a simplelinear model to look for linear association
cor.test(xbar,s2) # nearly the same as above,
#calculate the correlation and perform a test on corr=0
#(observe identical p-value to the lm for slope - why?)

#ii distr of xbar
mean(xbar)
sd(xbar)
2/sqrt(n)

hist(xbar,nclass=100,prob=TRUE)
xvec <- seq(min(xbar),max(xbar),length=100)
yvec <- dnorm(xvec,mu,sigma/sqrt(n))
lines(xvec,yvec,col=2)

#iii distr of s2
```

```
mean(s2*n/sigma^2)
var(s2*n/sigma^2)
# s2*n/sigma^2 should have
# mean n-1
#and variance 2*(n-1)
```

Problem 4: ANOVA

$$X_{ij} = \mu_i + \varepsilon_{ij}$$

where $i = 1, 2, 3$ refers to the three fund types, and $j = 1, \ldots, 6$ are the observations for each fund type. Further, ε_{ij} are assumed to be i.i.d. $N(0, \sigma^2)$.

```
ds=read.table("http://www.math.ntnu.no/~mettela/TMA4267/Data/funddat.txt",header=T)
dim(ds)
names(ds)

attach(ds)
# if you have y from Problem 2, do "rm(y)" before proceeding, if else,
# the new y in ds is masked by the old y.
```

a) Explain what this model means.

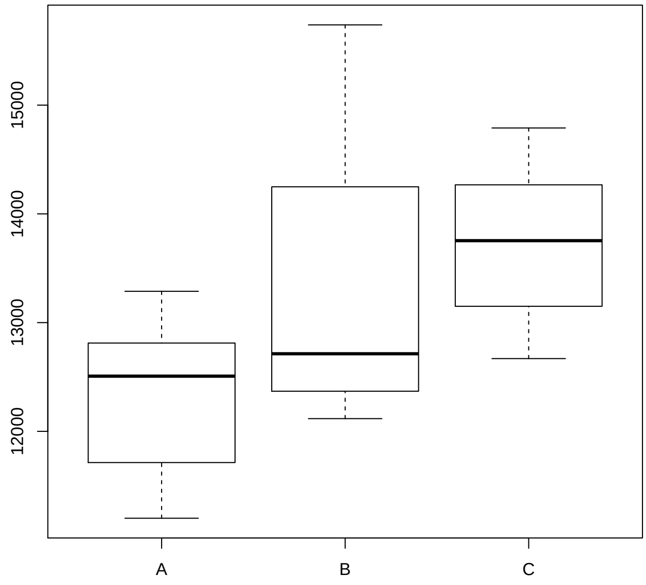
Each group has its own mean, and each group has the same variance. Observations are independent of each other.

```
boxplot(y~funds)
grandmean <- mean(y)
means <- ave(y,funds)
vars <- ave(y,funds,FUN=var)

means
 [1] 12338.17 12338.17 12338.17 12338.17 12338.17 12338.17 13316.67 13316.67
 [9] 13316.67 13316.67 13316.67 13316.67 13730.50 13730.50 13730.50 13730.50
[17] 13730.50 13730.50

vars
 [1]  603746.2  603746.2  603746.2  603746.2  603746.2  603746.2 1964107.5
 [8] 1964107.5 1964107.5 1964107.5 1964107.5 1964107.5  577199.5  577199.5
[15]  577199.5  577199.5  577199.5  577199.5
>
n <- length(y)
nis <- table(funds)
r <- length(nis)
```

Observe that `ave` produces one average corresponding to each observation, based on the group of the observation.



From the boxplot we observe that the inter quartile ranges are rather similar, but it is hard to conclude on any violations of assumptions.

b) We would like to test if the variance for fund type A and fund type B are the same.

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \text{ vs. } H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

Let S_i^2 be the sample variance for fund type i, $S_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2$ and $\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$.

From 1c) we know that if $V \sim \chi_p^2$ and $W \sim \chi_q^2$, where V and W are independent, then $F = \frac{V/p}{W/q}$ is Fisher distributed with p and q degrees of freedom. Further, we know that $(n_i - 1)S_i^2/\sigma_1^2$ is $\chi_{n_i-1}^2$, so that the ratio

$$\frac{\frac{(n_1-1)S_1^2/\sigma_1^2}{n_1-1}}{\frac{(n_2-1)S_2^2/\sigma_2^2}{n_2-1}} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

is Fisher with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. When the null hypothesis is true this ratio reduces to the suggested test statistics $\frac{S_1^2}{S_2^2}$. Remark: if we used the unbiased version of the sample variance (divide by n_i and not $n_i - 1$, then the suggested test statistics need in general to be scaled by $n_i/(n_i - 1)$ in the numerator and denominator to have a Fisher distribution.

We find $s_1^2 = 603746.2$, $s_2^2 = 1964107.5$, $n_1 = n_2 = 6$ and $f_{obs} = s_1^2/s_2^2 = 0.307$. This is a two-sided test so we reject if $f_{obs} < f_{\alpha/2}$ or $f_{obs} > f_{1-\alpha/2}$. In the F-distribution with $n_1 - 1 = 5$ and $n_2 - 1 = 5$ degrees of freedom the critical values are

```
> qf(0.025,5,5)
```

```
[1] 0.139931
> qf(0.975,5,5)
[1] 7.146382
```

This means that the null hypothesis is not rejected.

This is a two-sided test, and the p-value can be calculated as twice the smallest tail $2P(F_{5,5} < f_{obs})$, which give the p-value 0.22.

```
> 2*pf(vars[1]/vars[7],5,5)
[1] 0.221338
```

c) We now assume that $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma^2$, and perform a one-way ANOVA.

The one-way ANOVA hypothesis test is

$$\mu_1 = \mu_2 = \cdots = \mu_r \text{ vs. at least one pair differs}$$

The estimator for σ^2 is S^2 =MSE.

$$S^2 = \frac{\sum_{i=1}^{r}(n_i - 1)S_i^2}{n - r}$$

where $n = \sum_{i=1}^{r} n_i$.

```
grandmean <- mean(y)
means <- ave(y,funds)
n <- length(y)
nis <- table(funds)
r <- length(nis)

MS <- var(y)
SS <- MS*(n-1)

SSA <- sum((means-grandmean)^2)
MSA <- SSA/(r-1)

SSE <- sum((y-means)^2)
MSE <- SSE/(n-r)

F <- MSA/MSE
pf(F,r-1,n-r,lower.tail=FALSE)
[1] 0.08455223
```

Write down the ANOVA table (SSs, MSs, F and p-value).

Source	SS	df	MS	F	p-value
Treatment	6134625	2	3067312	2.93	0.085
Error	15725266	15	1048351		
Total	21859890	1285876			

d)

```
> res <- aov(y~as.factor(funds),data=ds)
> summary(res)
                 Df   Sum Sq Mean Sq F value Pr(>F)
as.factor(funds)  2  6134625 3067312   2.926 0.0846 .
Residuals        15 15725266 1048351
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

> names(res)
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "contrasts"     "xlevels"       "call"          "terms"
[13] "model"

> str(summary(res))
List of 1
 $ :Classes ’anova’ and ’data.frame’: 2 obs. of  5 variables:
  ..$ Df     : num [1:2] 2 15
  ..$ Sum Sq : num [1:2] 6134625 15725266
  ..$ Mean Sq: num [1:2] 3067312 1048351
  ..$ F value: num [1:2] 2.93 NA
  ..$ Pr(>F) : num [1:2] 0.0846 NA
 - attr(*, "class")= chr [1:2] "summary.aov" "listof"

> mode(summary(res))
[1] "list"

> summary(res)[[1]]
                 Df   Sum Sq Mean Sq F value  Pr(>F)
as.factor(funds)  2  6134625 3067312  2.9258 0.08455 .
Residuals        15 15725266 1048351
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

> sqrt(summary(res)[[1]][2,3])
[1] 1023.89
> sqrt(MSE)
[1] 1023.89
```

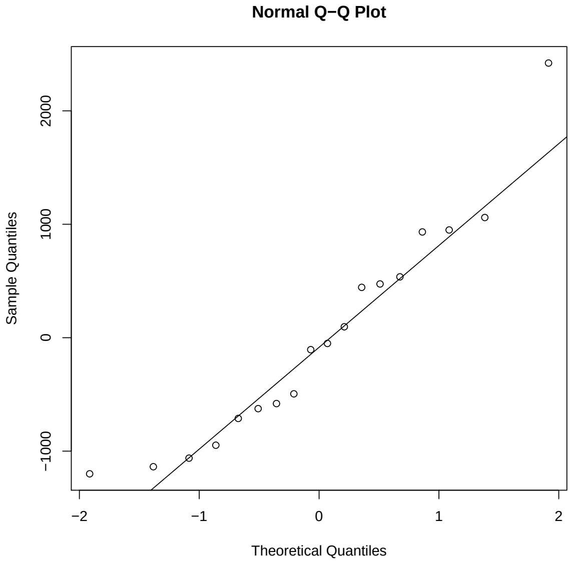
e) Normality of the error terms ε.

```
e <- res$residuals
ee <- y-means #alternative
sum(e-ee) #the same

qqnorm(e)
qqline(e)
library(nortest)
ad.test(e)
```

```
#better with studentized residuals, more later
es <- rstudent(res)
qqnorm(es)
qqline(es)
ad.test(es)
```

We see no reason to doubt the assumption of normality of errors.



f) Tests for homogeniety of variances: Bartlett and Levene.

```
> bartlett.test(y~as.factor(funds))

Bartlett test of homogeneity of variances

data:  y by as.factor(funds)
Bartlett's K-squared = 2.3914, df = 2, p-value = 0.3025
> library(car)
> leveneTest(y~as.factor(funds))
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  2  0.4712 0.6332
      15
```

Both tests reach the same conclusion: don't reject the null hypothesis of equal variances.

g)

$$H_0 : \mu_1 = (\mu_2 + \mu_3)/2 \text{ vs. } H_1 : \mu_1 \neq (\mu_2 + \mu_3)/2$$

$U = \bar{X}_1 - \frac{\bar{X}_2+\bar{X}_3}{2}$, where $\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$, n_i is the number of observations for group i. U is a linear combination of normal RVs and thus U is normal. We calculate mean and variance when the null hypothesis is true. We also assume $n_1 = n_2 = n_3 = n_* = 6$.

$$\begin{aligned}
\mathrm{E}(U) &= \mathrm{E}(\bar{X}_1) - \frac{1}{2}(\mathrm{E}(\bar{X}_2) + \mathrm{E}(\bar{X}_3)) \\
&= \mu_1 - \frac{1}{2}(\mu_2 + \mu_3) = 0 \\
\mathrm{Var}(U) &= \mathrm{Var}(\bar{X}_1) + (\frac{1}{2})^2(\mathrm{Var}(\bar{X}_2) + \mathrm{Var}(\bar{X}_3)) \\
&= \sigma^2/n_1 + (\sigma^2/n_2 + \sigma^2/n_3)/4 \\
&= (\sigma^2/n_*) \cdot (1 + 2/4) \\
&= (\sigma^2/6) \cdot (6/4) = \sigma^2/4
\end{aligned}$$

That is, $U \sim N(0, \sigma^2/4)$. A natural test statistic will be

$$T\frac{U}{S/\sqrt{4}} = \frac{2U}{S}$$

Since $2U/\sigma$ is standard normal and $(n-3)S^2/\sigma^2$ is χ^2_{n-3}, then T must be t-distributed with $n - r = 18 - 3 = 15$ degrees of freedom.

```
> U <- means[1]-(means[7]+means[13])/2
> U
[1] -1185.417
> S <- sqrt(MSE)
> T <- 2*U/S
> T
[1] -2.315515
> 2*pt(T,n-r)
[1] 0.03515318
```

The p-value is 0.035, reject the null hypothesis.