Homework 9

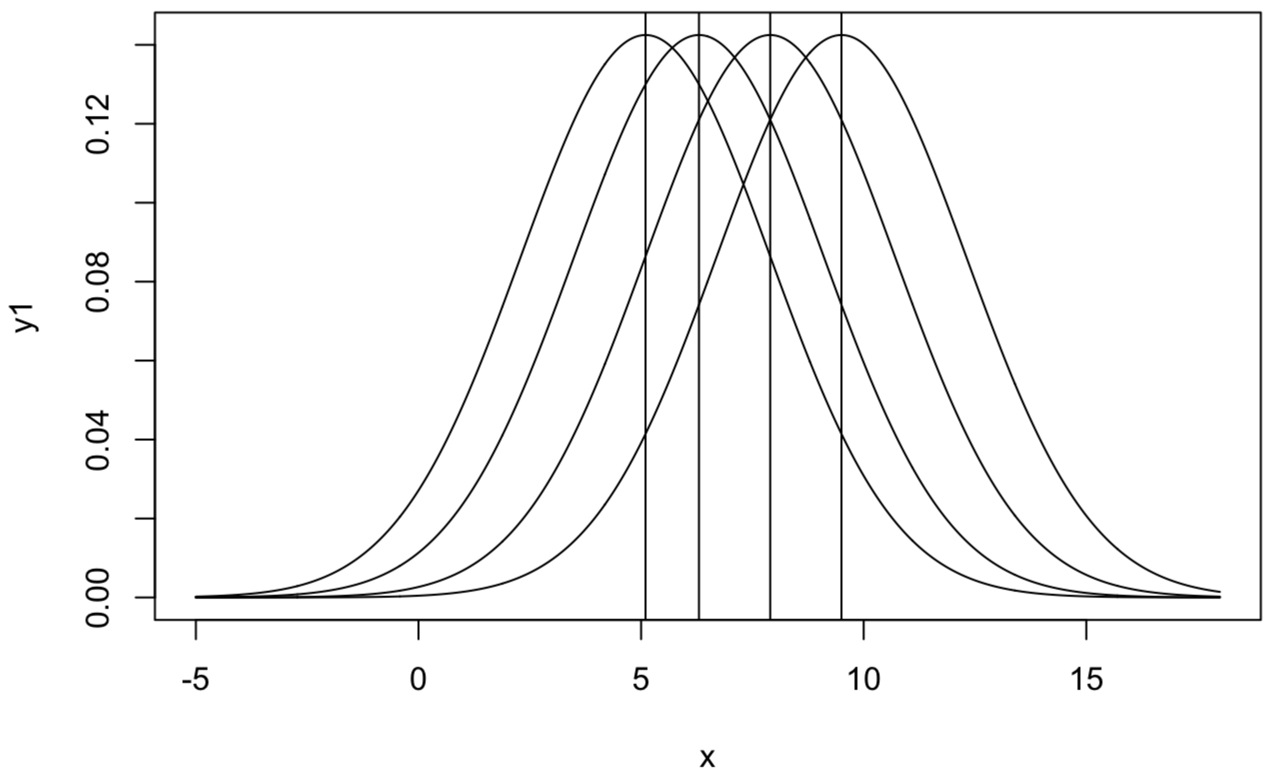
CHEN XUPENG

2018/5/20

1

a

```
x <- seq(-5, 18, length=1000)
y1 <- dnorm(x, mean=5.1, sd=2.8)
y2 <- dnorm(x, mean=6.3, sd=2.8)
y3 <- dnorm(x, mean=7.9, sd=2.8)
y4 <- dnorm(x, mean=9.5, sd=2.8)
plot(x, y1, type="l", lwd=1)
lines(x, y2, type="l", lwd=1)
lines(x, y3, type="l", lwd=1)
lines(x, y4, type="l", lwd=1)
abline(v=5.1)
abline(v=6.3)
abline(v=7.9)
abline(v=9.5)
```



b

$$E(MSE) = \sigma^2 = 7.84$$

$$E(MSTR) = \sigma^2 + \frac{\sum_i (\mu_i - \mu)^2}{\gamma - 1}$$

$$= 7.84 + \frac{100[(5.1-7.2)^2 + (6.3-7.2)^2 + (7.9-7.2)^2 + (9.5-7.2)^2]}{3} \approx 374$$

It suggests that the different treatments have substantially impact on Y

c

$$\text{Use same equation as b, we have: } E(MSTR) = \sigma^2 + \frac{\sum_i (\mu_i - \mu)^2}{\gamma - 1}$$

$$= 7.84 + \frac{100[(5.1-7.2)^2 + (5.6-7.2)^2 + (9-7.2)^2 + (9.5-7.2)^2]}{3} \approx 523$$

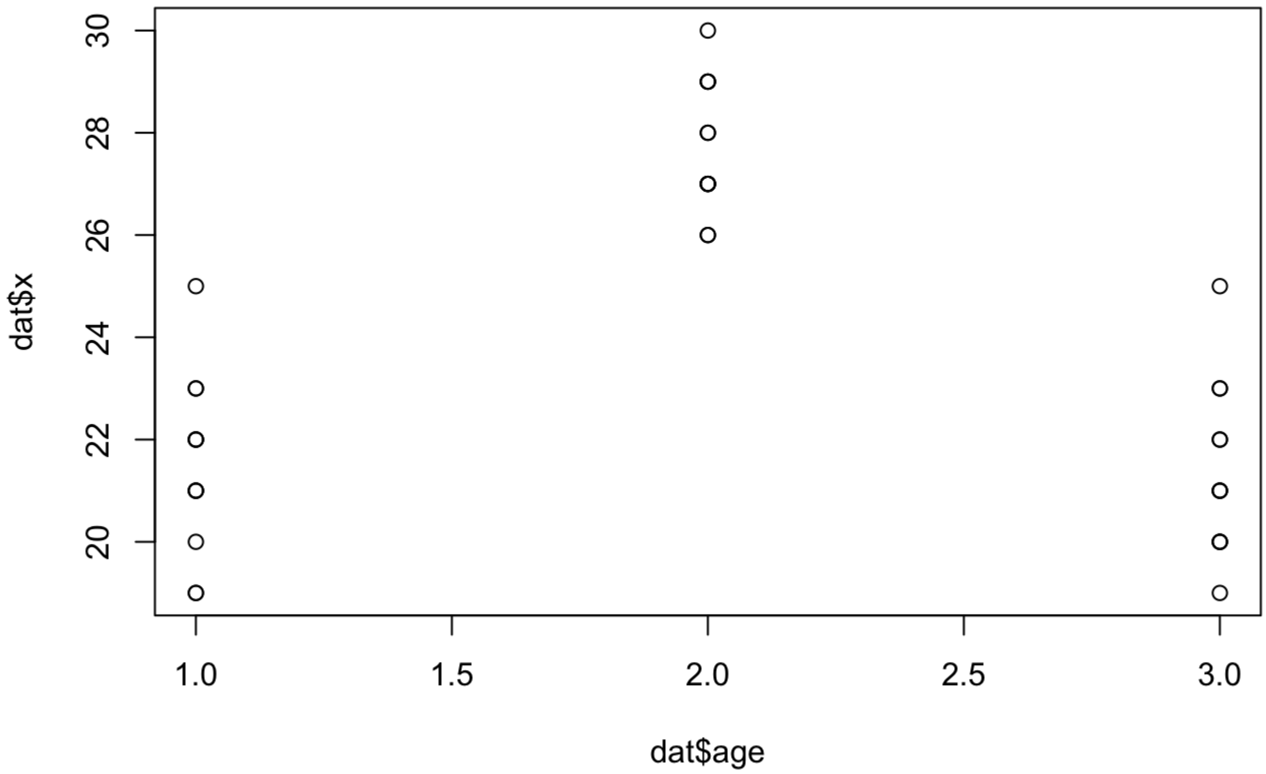
It is because the points distribution are more scattered compared to b.

2

```
dat<-read.table("CH16PR10.txt")
names(dat)<-c("x","age","num")
```

a

```
plot(x=dat$age, y=dat$x)
```



The factor level means seem to differ, at least the middle ge group differs from the other two. The variability within each factor level seems to be constant.

b

```
code <- rbind(diag(1,3),-1)
xx <- code[dat$age,]
dat$x1 <- xx[,1]
dat$x2 <- xx[,2]
dat$x3 <- xx[,3]

dat$age <- factor(dat$age)
fit1 <- aov(x~age,data=dat)
yhat <- fitted(fit1)
yhat
```

```
##        1        2        3        4        5        6        7        8
## 21.50000 21.50000 21.50000 21.50000 21.50000 21.50000 21.50000 21.50000
##        9       10       11       12       13       14       15       16
## 21.50000 21.50000 21.50000 21.50000 27.75000 27.75000 27.75000 27.75000
##       17       18       19       20       21       22       23       24
## 27.75000 27.75000 27.75000 27.75000 27.75000 27.75000 27.75000 27.75000
##       25       26       27       28       29       30       31       32
## 21.41667 21.41667 21.41667 21.41667 21.41667 21.41667 21.41667 21.41667
##       33       34       35       36
## 21.41667 21.41667 21.41667 21.41667
```

c

```
resid1 <- resid(fit1)
resid1
```

```
##          1          2          3          4          5          6 
##  1.5000000  3.5000000 -0.5000000  0.5000000 -0.5000000  0.5000000 
##          7          8          9         10         11         12 
## -1.5000000  1.5000000 -2.5000000  0.5000000 -2.5000000 -0.5000000 
##         13         14         15         16         17         18 
##  0.2500000 -0.7500000 -0.7500000  1.2500000 -1.7500000  1.2500000 
##         19         20         21         22         23         24 
## -0.7500000  2.2500000  0.2500000 -0.7500000 -1.7500000  1.2500000 
##         25         26         27         28         29         30 
##  1.5833333 -1.4166667  3.5833333 -0.4166667  0.5833333  1.5833333 
##         31         32         33         34         35         36 
## -0.4166667 -1.4166667 -2.4166667 -1.4166667  0.5833333 -0.4166667
```

d

```
summary(fit1)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## age          2  316.7  158.36    63.6 4.77e-12 ***
## Residuals   33   82.2    2.49                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

e

$$H_0 : \mu_1 = \mu_2 = \mu_3$$
$$H_1 : \text{otherwise}$$
$$F^* = \frac{MSTR}{MSE}$$

Reject null hypothesis if F*>$F_{0.99,3,33}$ The p value is 4.769e-12 We reject H_0

f

If seems that middle aged people tend to offer more cash for a used car, while young and old people tend to offer less.

3

a

```
fit2 <- lm(x~age,data=dat)
summary(fit2)
```

```
##
## Call:
## lm(formula = x ~ age, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5000 -0.9167 -0.4167  1.2500  3.5833
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.50000    0.45551  47.200  < 2e-16 ***
## age2         6.25000    0.64419   9.702 3.43e-11 ***
## age3        -0.08333    0.64419  -0.129    0.898
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.578 on 33 degrees of freedom
## Multiple R-squared:  0.794,  Adjusted R-squared:  0.7815
## F-statistic:  63.6 on 2 and 33 DF,  p-value: 4.769e-12
```

The model is $Yhat = 21.5 + 6.25X_1 - 0.0833X_2$ The intercept term estimates the average cell sample mean.

b

```
anova(fit2)
```

```
## Analysis of Variance Table
##
## Response: x
##           Df Sum Sq Mean Sq F value    Pr(>F)
## age        2 316.72  158.36  63.601 4.769e-12 ***
## Residuals 33  82.17    2.49
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H_0 : \tau_1 = \tau_2$

$H_1 : otherwise$

$F^* = \frac{MSR}{MSE}$

Reject null hypothesis if F*> $F_{0.99,2,33}$ The p value is 4.769e-12, so We reject H_0

4

b

```
Young=c(23, 25, 21, 22, 21, 22, 20, 23, 19, 22, 19, 21)
Middle=c(28, 27, 27, 29, 26, 29, 27, 30, 28, 27, 26, 29)
Elderly=c(23, 20, 25, 21, 22, 23, 21, 20, 19, 20, 22, 21)
FactorLevels=c(1,2,3)
n1=length(Young)
n2=length(Middle)
n3=length(Elderly)

MyData=data.frame(
Values=c(Young,Middle,Elderly),
Treatment=c(rep(1,n1),rep(2,n2),rep(3,n3)))

y=MyData$Values
x=factor(MyData$Treatment)
means=tapply(y,x,mean)
n=tapply(y,x,length)
df=sum(n)-2

MSE=2.49
alpha=0.01
l1=means[1]-qt(1-alpha/2,df)*sqrt(MSE/n[1])
u1=means[1]+qt(1-alpha/2,df)*sqrt(MSE/n[1])
print(c(l1,u1))
```

```
##        1        1
## 20.25716 22.74284
```

So the confidence level is (20.2572,22.7428)

C

```
MSE=2.49
alpha=0.01
l31=means[3]-means[1]-qt(1-alpha/2,df)*sqrt(MSE/n[1]+MSE/n[3])
u31=means[3]-means[1]+qt(1-alpha/2,df)*sqrt(MSE/n[1]+MSE/n[3])
print(c(l31,u31))
```

```
##         3         3
## -1.840978  1.674312
```

This confidence interval contains 0, so we cannot reject the null hypothesis that $\mu_1 = \mu_3$ ## d

```
MSE=2.49;
alpha=0.01;
lcontrast=-means[1]+2*means[2]-means[3]-qt(1-alpha/2,df)*sqrt(MSE/n[1]+4*MSE/n[2]+MSE
/n[3])
ucontrast=-means[1]+2*means[2]-means[3]+qt(1-alpha/2,df)*sqrt(MSE/n[1]+4*MSE/n[2]+MSE
/n[3])
print(c(lcontrast,ucontrast))
```

```
##         1         1
##  9.539003 15.627664
```

$H_0 : \mu_2 - \mu_1 = \mu_3 - \mu_2$ $H_1 : otherwise$ Since the confidence interval for the contrast does not contain 0, we do not reject the null hypothesis.

e

```
results<-aov(y~x);
TukeyHSD(results)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = y ~ x)
##
## $x
##            diff       lwr       upr     p adj
## 2-1  6.25000000  4.669286  7.830714 0.0000000
## 3-1 -0.08333333 -1.664048  1.497381 0.9908192
## 3-2 -6.33333333 -7.914048 -4.752619 0.0000000
```

There is significant difference between young and middle aged people, as well as mlderly and middle aged people. But there is no significant difference between the young and elderly.

f

```
pairwise.t.test(y,x,p.adjust="bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  y and x
##
##   1       2
## 2 1.0e-10 -
## 3 1       7.4e-11
##
## P value adjustment method: bonferroni
```

The bonferroni method gives the same result. But it won't be more efficient, since it "overstates" the significnce level.

5

a

This has been done in the previous problem, question d. The confidence interval is (-15.627664,-9.539003)

b

D_1 = 6.2500, D_2 = −6.3333, D_3 = −0.0833,

L_1 =−12.5833, s{D_i} =0.6442 (i = 1, 2, 3), s{L_1} = 1.1158,

$F(0.90, 2, 33) = 2.47$, $S = 2.223$

Then we can obtain the family intervals:

$(4.818, 7.682)$

$(-7.765, -4.901)$

$(-1.515, 1.349)$

$(-15.064, -10.103)$