STATS 3N03/3J04
ASSIGNMENT #3 – SOLUTIONS
2005-12-06
Corrected 2005-12-14

[Full Marks = 140]

Question 1 [10]

Figure 8-4

```
> xgr <- seq(-4,4,length=50)
> plot(xgr, dnorm(xgr), type = "l", lty = 1, xlab = "x", ylab ="f(x)")
> lines(xgr,dt(xgr,10),lty=2)
> lines(xgr,dt(xgr,1),lty=3)
> legend(1.8,.38,c("infinite df","10 df","1 df"),lty=1:3)
> title("t density")
```

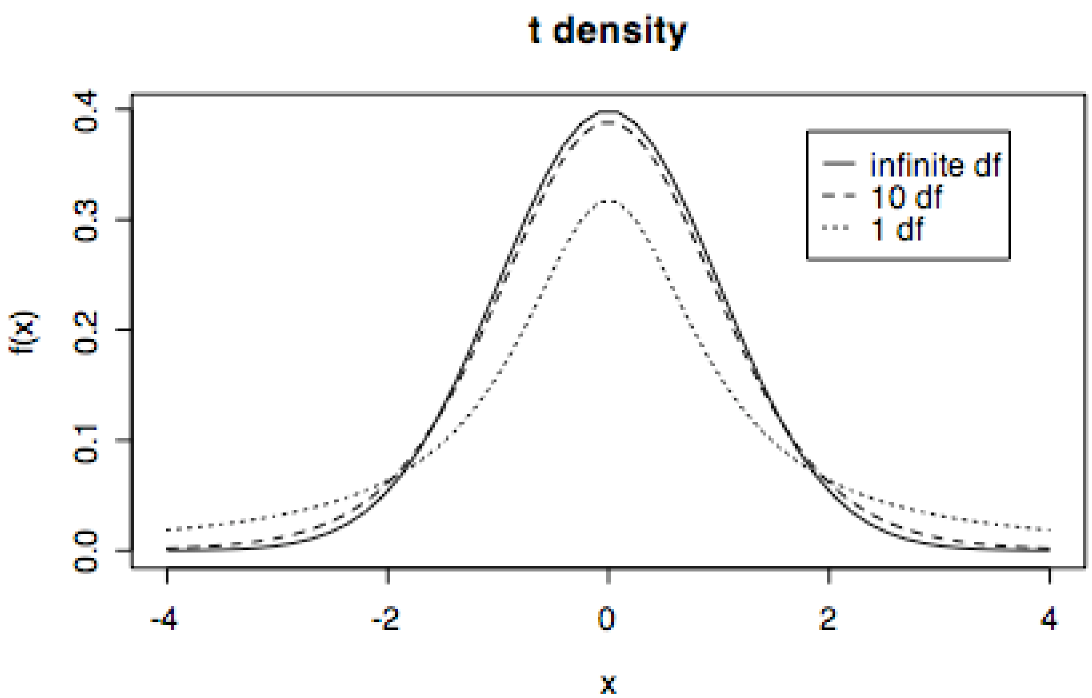


Figure 8-8

```
> xgr <- seq(0,30,length=50)
> plot(xgr, dchisq(xgr, 2), type = "l", lty = 1, xlab = "x", ylab ="f(x)")
> lines(xgr,dchisq(xgr,5),lty=2)
> lines(xgr,dchisq(xgr,10),lty=3)
> legend(15,.4,c("2 df","5 df","10 df"),lty=1:3)
> title("Chi-square density")
```

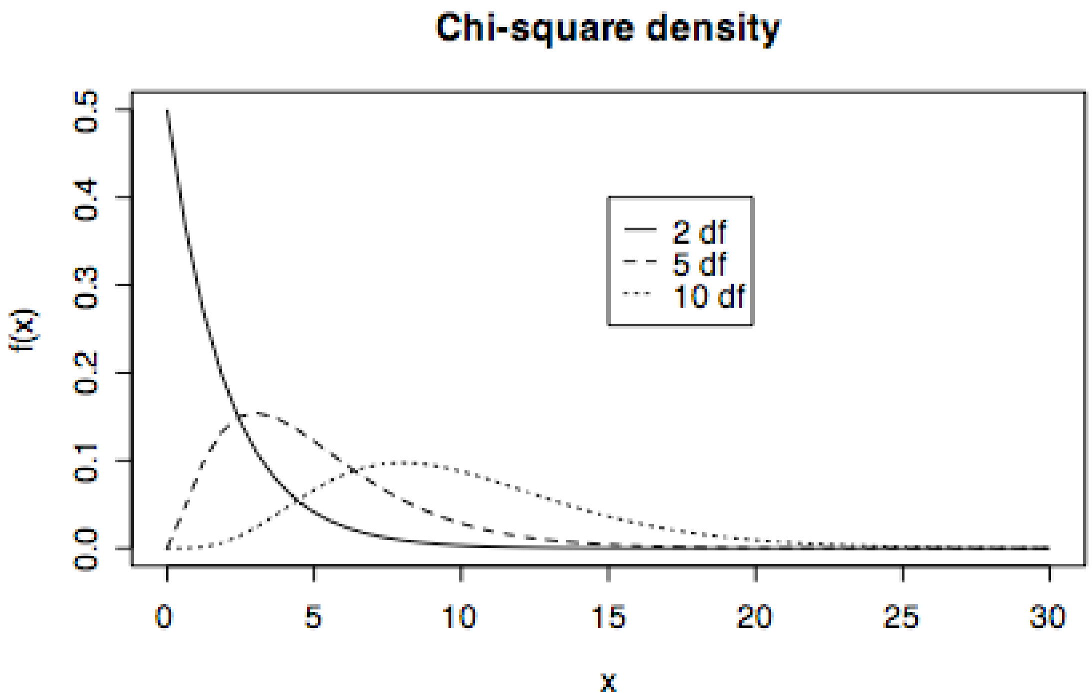
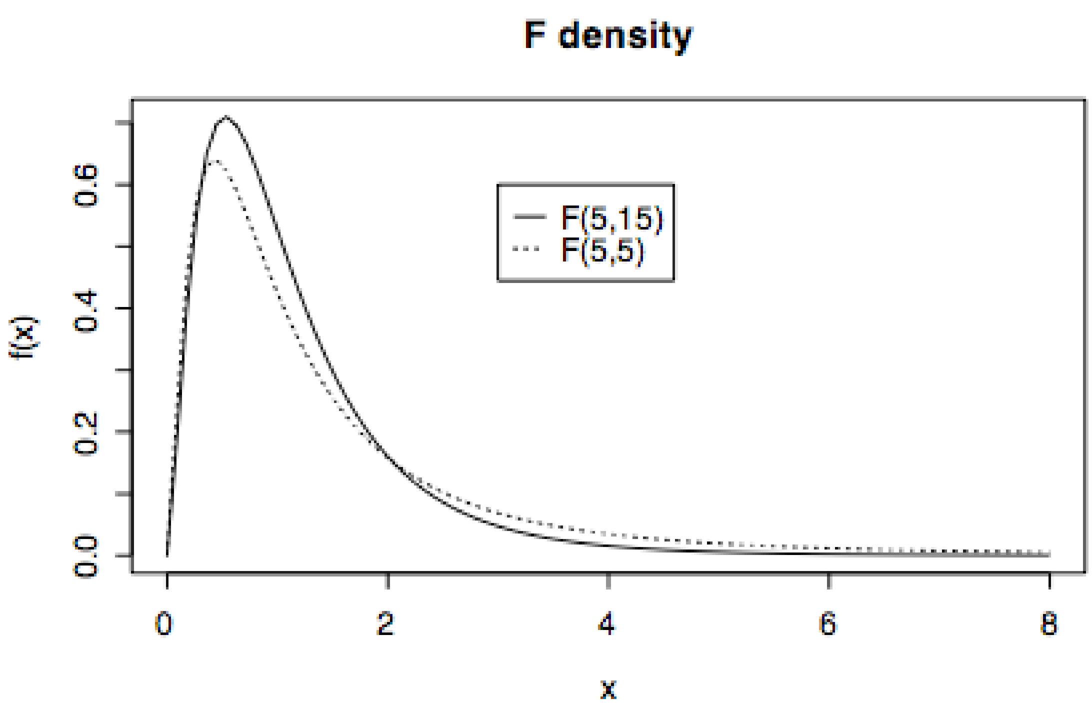


Figure 10

```
> xgr <- seq(0,8,length=90)
> plot(xgr, df(xgr,5,15), type = "l", lty = 1, xlab = "x", ylab ="f(x)")
> lines(xgr,df(xgr,5,5),lty=3)
> legend(3,.6,c("F(5,15)","F(5,5)"),lty=c(1,3))
> title("F density")
```



Question 2 [10]

When n = 4, the coverage seems to be slightly less than 95%, closer to 93% or 94%. With such a small difference, 1000 simulated intervals aren't enough to answer the question. However, it would be safe to say that n = 20 is enough. I wrote a function so I wouldn't have to keep re-entering the code to try more examples.

```
> weibconf
function (n, shape, scale, nint = 1000)
{
    wmean <- scale * gamma(1 + 1/shape)
    weibdata <- matrix(rweibull(nint * n, shape, scale), ncol = n)
    xbar <- apply(weibdata, 1, mean)
    sx <- apply(weibdata, 1, sd)
    llim <- xbar - qt(0.975, n - 1) * sx/sqrt(n)
    ulim <- xbar + qt(0.975, n - 1) * sx/sqrt(n)
    mean(wmean > llim & wmean < ulim)
}
> weibconf(4, 30, 2)
[1] 0.932
> weibconf(4, 30, 2)
[1] 0.938
> weibconf(20, 30, 2)
[1] 0.956
> weibconf(20, 30, 2)
[1] 0.934
> weibconf(100, 30, 2, 100000)
[1] 0.94778
```

Question 3 **[4 + 4 + 4 + 8]**

(a) You need at least 29 degrees of freedom.

```
> for (df in 25:30) print(c(df,qchisq(.995,df)/qchisq(.005,df)))
[1] 25.000000  4.460974
[1] 26.000000  4.326958
[1] 27.000000  4.204493
[1] 28.000000  4.092128
[1] 29.000000  3.988646
[1] 30.000000  3.893019
```

(b) Here, $\alpha = 0.01$, $\beta = 0.10$, $\delta = 0.5$ and $\sigma = 1.3$, so the required sample size n is given by

```
> ((qnorm(1-(0.01)/2) + qnorm(1-0.1))*1.3/0.5)^2
[1] 100.5847
```

or, using table values from the text

```
> ((2.576 + 1.282)*1.3/0.5)^2
[1] 100.6169
```

so 101 observations would be required.

The probability of a Type II error is computed by text formula (9-17); when $n = 10$ it gives

```
> pnorm(qnorm(1-(0.01)/2)-0.5*sqrt(10)/1.3) + pnorm(-qnorm(1-
(0.01)/2)-0.5*sqrt(10)/1.3)
[1] 0.9130915
```

and this probability is much too high for the test to be useful.

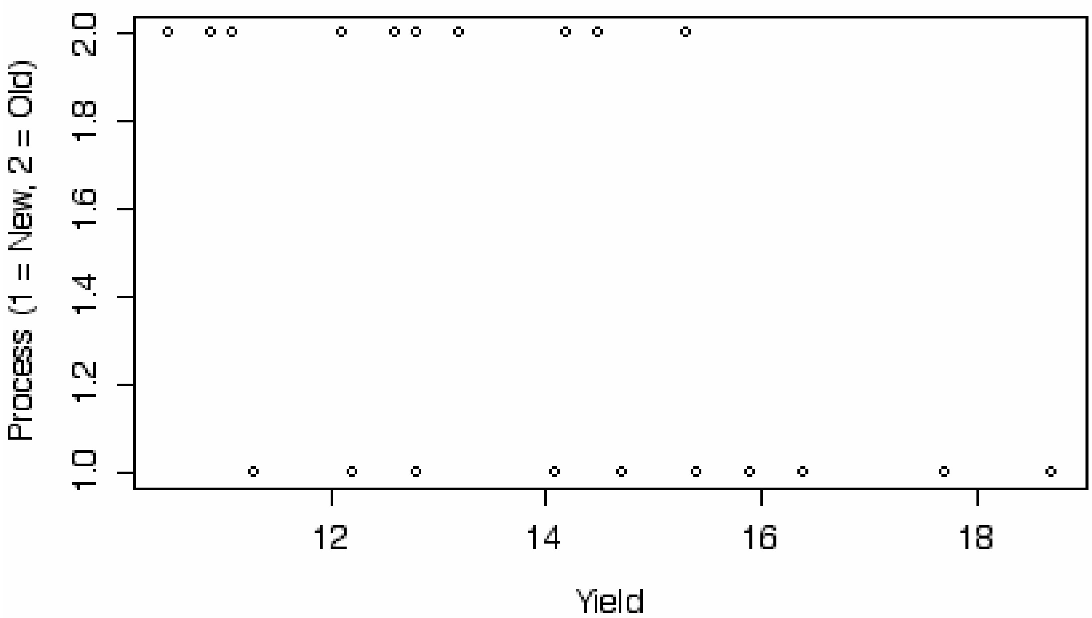
(c) Let D be the event that the lot was produced domestically and let D' be the event that it was produced offshore. Let X be the number of defective items in a lot of 100. We are given that P(D) = 0.1, P(D') = 0.9. Assuming independence of defective items, we have that X | D ~ Bin(100, 0.02) and X | D' ~ Bin(100, 0.01). Hence, by Bayes' theorem,

P(D | X = 3) = P(X = 3 | D)*P(D)/(P(X = 3 | D)*P(D) + P(X = 3 | D')*P(D')) = 0.249.

```
> dbinom(3,100,0.02)*0.1/
(dbinom(3,100,0.02)*0.1 + dbinom(3,100,0.01)*0.9)
[1] 0.2492600
```

Question 4 [15 + 10]

(a) The correct analysis is an independent-sample t-test to compare the means, assuming homoscedasticity. The graph could be comparative dot plots, box plots, stem and leaf plots, or histograms, but they must be comparative (side by side, or one above the other, on identical scales).




```
> t.test(yield~process, coal, var.equal=T)

     Two Sample t-test

data:  yield by process
t = 2.4159, df = 18, p-value = 0.02654
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 0.2868403 4.1131597
sample estimates:
mean in group New mean in group Old
            14.92             12.72
```

Testing for homoscedasticity:

```
> var(coal$yield[coal$process=="New"])
[1] 5.679556
> var(coal$yield[coal$process=="Old"])
[1] 2.612889
> var(coal$yield[coal$process=="New"])/
var(coal$yield[coal$process=="Old"])
[1] 2.173669
> 2*(1-pf(var(coal$yield[coal$process=="New"])
/var(coal$yield[coal$process=="Old"]),9,9))
```

```
[1] 0.2629612
```

A two-sided F test on 9 over 9 df gives P > 0.1, so there is no evidence from these data of heteroscedasticity.

Testing equality of the means without assuming homoscedasticity, we get an almost identical result:

```
> t.test(yield~process, coal)

     Welch Two Sample t-test

data:  yield by process
t = 2.4159, df = 15.834, p-value = 0.02816
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 0.2679119 4.1320881
sample estimates:
mean in group New mean in group Old
            14.92              12.72
```

Additional Assumptions: Normality (looks OK in dot plot), Independence (small sample, can't test).

Conclusions: There is no evidence (P > 0.1) of heteroscedasticity. There is some evidence (0.05 > P > 0.025 two –sided, 0.025 > P > 0.01 right-tailed) that the means are not the same, so we conclude that the new process gives a slightly higher yield than the old process.

(b) The correct analysis is a paired t-test. The graph could be a dot plot, stem and leaf plot, box plot or histogram of the differences.

A regression analysis with a test of the slope is not appropriate as it would say nothing about the difference in heat loss between glass and steel pipes, only about their similarity at different diameters.

```
> heatloss
  steel glass diff
1   4.6   2.5  2.1
2   3.7   1.3  2.4
3   4.2   2.0  2.2
4   1.9   1.8  0.1
5   4.8   2.7  2.1
6   6.1   3.2  2.9
7   4.7   3.0  1.7
8   5.5   3.5  2.0
9   5.4   3.4  2.0


> stem(heatloss$diff)
```

```
  The decimal point is at the |

  0 | 1
  1 | 7
  2 | 0011249

> mean(heatloss$diff)/sqrt(var(heatloss$diff)/9)
[1] 7.608696
> 1-pt(mean(heatloss$diff)/sqrt(var(heatloss$diff)/9), 8)
[1] 3.126906e-05
> 2*(1-pt(mean(heatloss$diff)/sqrt(var(heatloss$diff)/9), 8))
[1] 6.253811e-05

> t.test(heatloss$steel,heatloss$glass,pair=T)

     Paired t-test

data:  heatloss$steel and heatloss$glass
t = 7.6087, df = 8, p-value = 6.254e-05
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 1.355132 2.533757
sample estimates:
mean of the differences
               1.944444
```

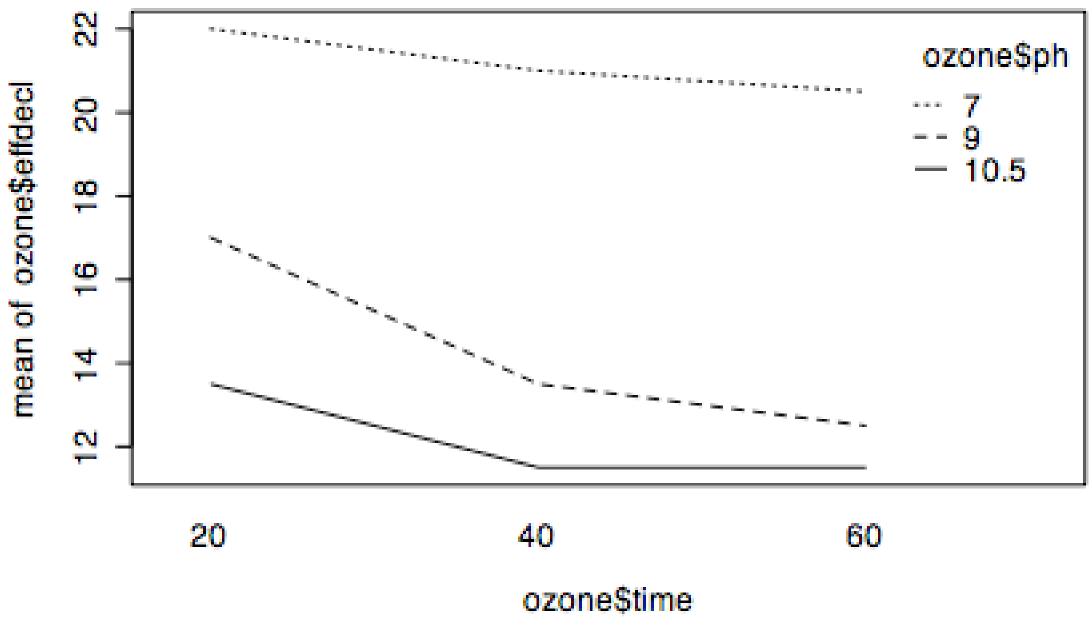
The t-test could be either right-tail or two-tail but either way P << 0.001 so there is strong evidence from these data that heat loss in glass pipes is less than in steel pipes.

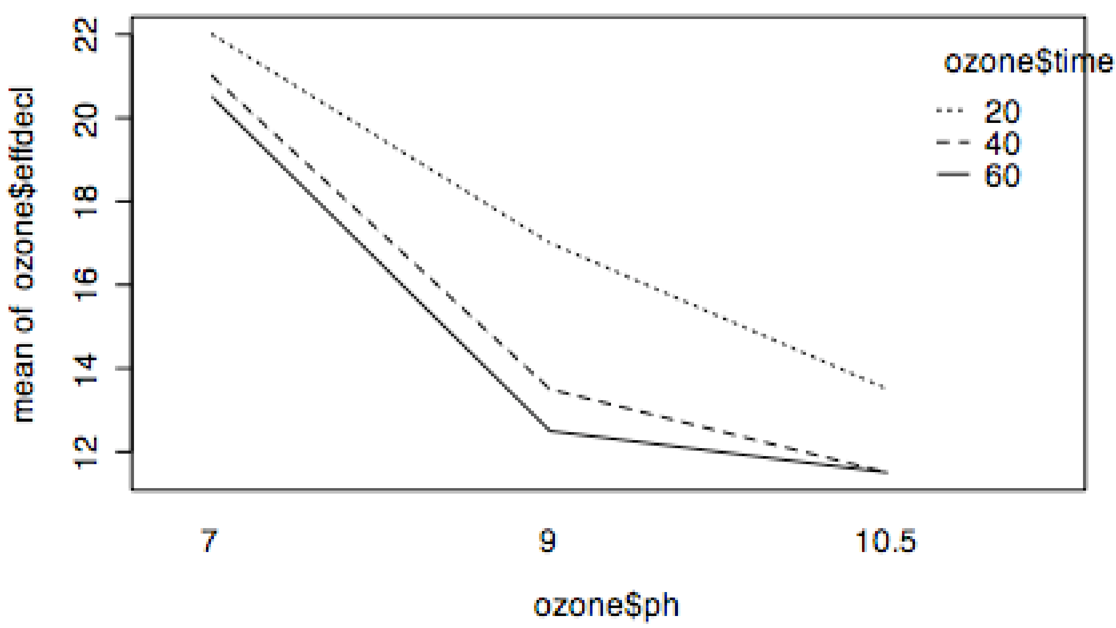
Assumptions: The differences are independent (sample size is too small to test) and normal (sample size is too small to test).

Conclusions: There is strong evidence (P << 0.001 by a one-sided or two-sided test) that heat loss in glass pipes is less than in steel pipes.

Question 5 [25]

```
> interaction.plot(ozone$time, ozone$ph, ozone$effdecl)
> interaction.plot(ozone$ph, ozone$time, ozone$effdecl)
```



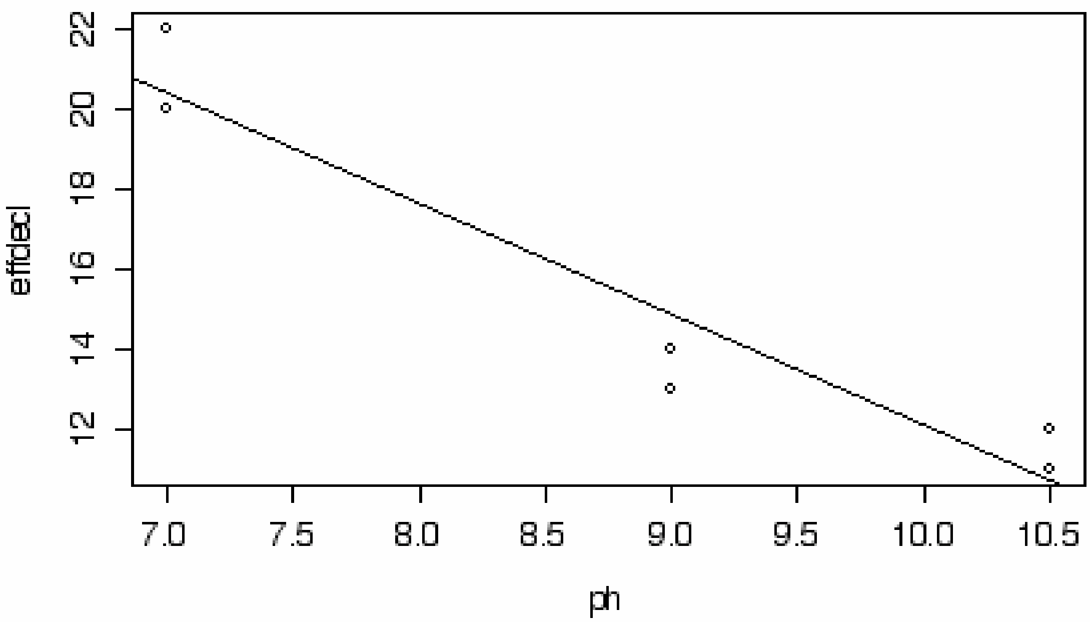


```
> ozone
   effdecl time   ph
1       23   20  7.0
2       21   20  7.0
3       16   20  9.0
4       18   20  9.0
5       14   20 10.5
6       13   20 10.5
7       20   40  7.0
8       22   40  7.0
9       14   40  9.0
10      13   40  9.0
11      12   40 10.5
12      11   40 10.5
13      21   60  7.0
14      20   60  7.0
15      13   60  9.0
16      12   60  9.0
17      10   60 10.5
18      13   60 10.5

> anova(lm(effdecl~as.factor(time)*as.factor(ph), ozone))
Analysis of Variance Table

Response: effdecl
                             Df  Sum Sq Mean Sq F value    Pr(>F)    
as.factor(time)               2  24.111  12.056  8.3462  0.008912 ** 
as.factor(ph)                 2 264.778 132.389 91.6538 1.038e-06 ***
as.factor(time):as.factor(ph) 4   5.889   1.472  1.0192  0.447259    
Residuals                     9  13.000   1.444                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> anova(lm(effdecl~as.factor(time)*as.factor(ph), ozone))[4,3]
[1] 1.444444

> 9*anova(lm(effdecl~as.factor(time)*as.factor(ph), 
ozone))[4,3]/c(qchisq(.975,9), qchisq(.025,9))
[1] 0.6833916 4.8141203
```

Assumptions: Normality, Independence, Homoscedasticity.

Conclusions: There is no evidence ($P > 0.1$) of an interaction between reaction time and pH level, so we can test the main effects. There is strong evidence that both time ($P << 0.01$) and pH level ($P << 0.01$) affect the mean percent decline in effluent.

Question 6 [25]




```
> plot(effdecl~ph, ozone[ozone$time==40,])
> abline(lm(effdecl~ph, ozone[ozone$time==40,]))


> anova(lm(effdecl~ph, ozone[ozone$time==40,]))
Analysis of Variance Table

Response: effdecl
          Df Sum Sq Mean Sq F value   Pr(>F)
ph         1 94.651  94.651  43.606 0.002725 **
Residuals  4  8.682   2.171
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1


> anova(lm(effdecl~ph+as.factor(ph), ozone[ozone$time==40,]))
Analysis of Variance Table

Response: effdecl
              Df Sum Sq Mean Sq F value   Pr(>F)
ph             1 94.651  94.651 94.6509 0.002307 **
as.factor(ph)  1  5.682   5.682  5.6824 0.097285 .
Residuals      3  3.000   1.000
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

Using the regression residual:

```
> anova(lm(effdecl~ph,
ozone[ozone$time==40,]))["Residuals","Mean Sq"]

2.170608
> 4*anova(lm(effdecl~ph,
ozone[ozone$time==40,]))["Residuals","Mean Sq"]
/c(qchisq(.975,4),qchisq(.025,4))
[1]  0.7791626 17.9234100
```

Using pure error:

```
> anova(lm(effdecl~ph+as.factor(ph),
ozone[ozone$time==40,]))["Residuals","Mean Sq"]

1
> 3*anova(lm(effdecl~ph+as.factor(ph),
ozone[ozone$time==40,]))["Residuals","Mean Sq"]
/c(qchisq(.975,3),qchisq(.025,3))
[1]  0.3209104 13.9020648
```

Assumptions: Linear relationship (OK by lack of fit test), Independence (can't test), Homoscedasticity (looks OK on plot).

Conclusions: There is no evidence from these data ($P = 0.1$) that the relationship between percent decline in effluent and pH is not linear over the range of pH studied, at 40 min reaction times. There is strong evidence ($P << 0.01$ using either the regression residual or pure error) that the slope of the relationship is not zero.

```
> predict(lm(effdecl~ph,
ozone[ozone$time==40,]),newdat=data.frame(ph=8))
[1] 17.64189
```

By interpolation of the fitted line, we predict a 17.6% decline in effluent when pH = 8. Since this is an interpolation of a relationship demonstrated to be linear, it can be considered reliable.

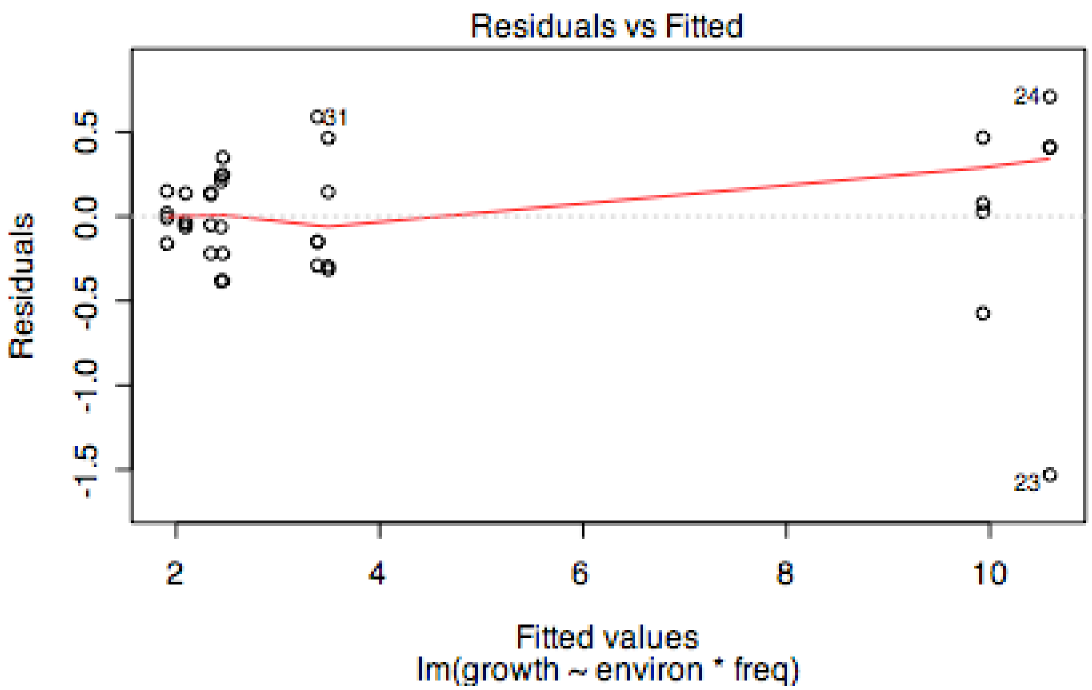
Question 7 [25]

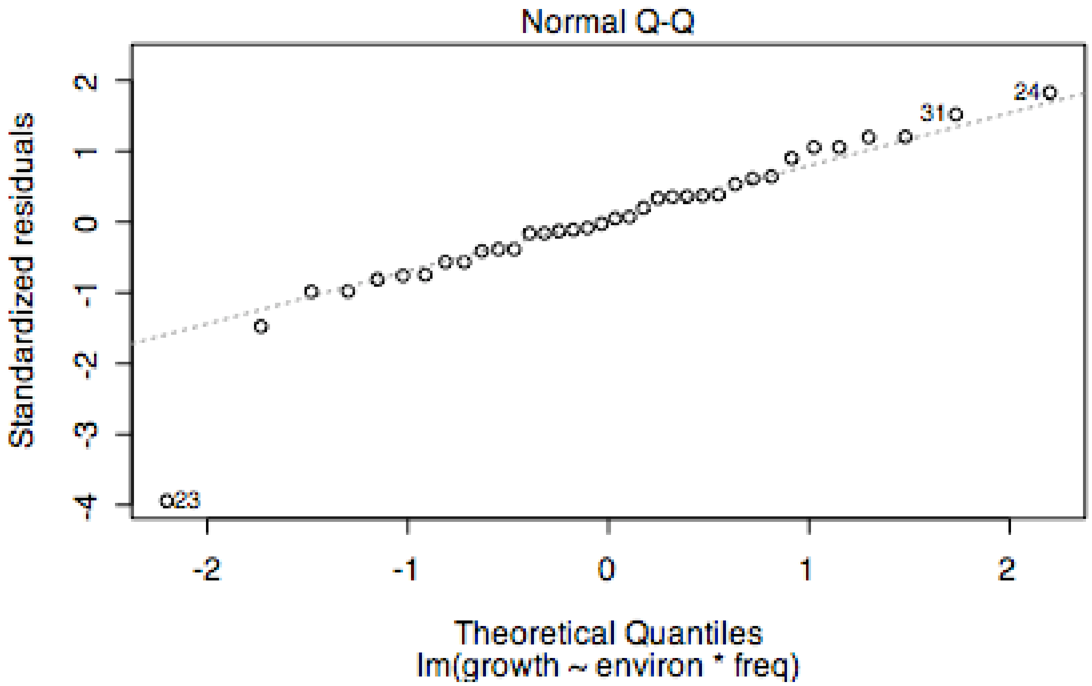
The analyses in original units and on a log scale give very similar results and lead to the same conclusion: the interaction is significant at the 5% level (or, better to say, $P \ll 0.001$ so there is very strong evidence of an interaction between frequency and environment). That means that both frequency and environment affect the crack growth rate, but the effect of the environment is different at different frequencies; the higher the frequency, the less difference the environment makes. Because the interaction is significant, we do not test the main effects.

The residual plots show that the residuals from the log-scale analysis follow a normal distribution more closely than residuals from the original-scale analysis. In the original scale, the 23rd observation is an outlier with a large negative residual.

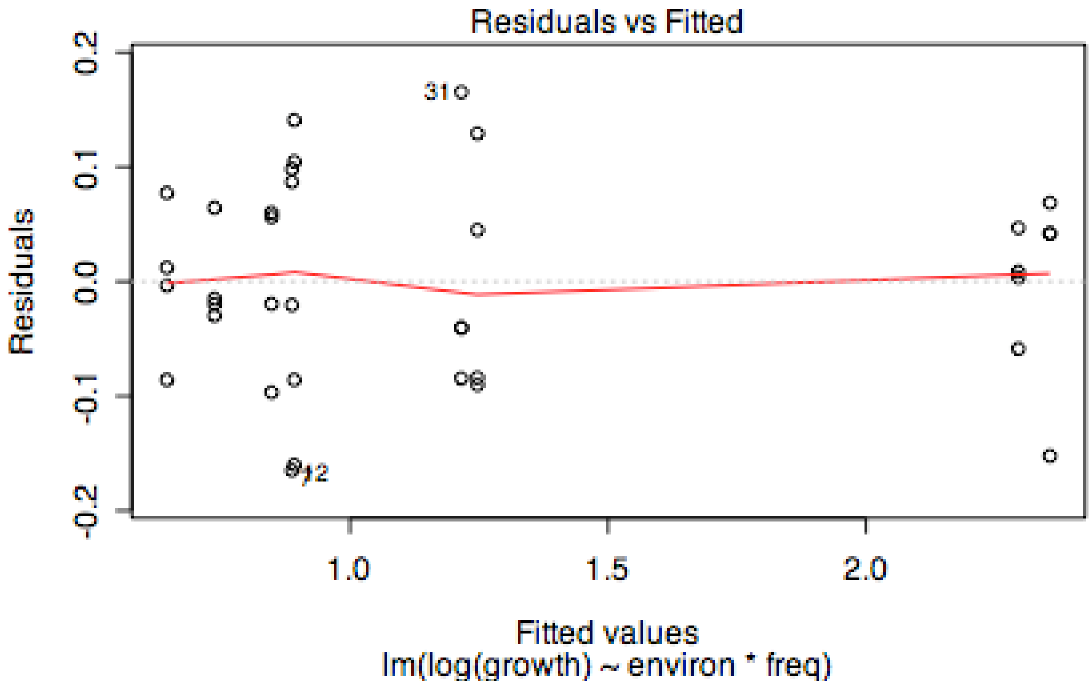
```
> cracks
   growth   environ freq
1    2.29       Air   10
2    2.47       Air   10
3    2.48       Air   10
4    2.12       Air   10
5    2.65       Air    1
6    2.68       Air    1
7    2.06       Air    1
8    2.38       Air    1
9    2.24       Air  0.1
10   2.71       Air  0.1
11   2.81       Air  0.1
12   2.08       Air  0.1
13   2.06     Water   10
14   2.05     Water   10
15   2.23     Water   10
16   2.03     Water   10
17   3.20     Water    1
18   3.18     Water    1
19   3.96     Water    1
20   3.64     Water    1
21  11.00     Water  0.1
22  11.00     Water  0.1
23   9.06     Water  0.1
24  11.30     Water  0.1
25   1.90 Saltwater   10
26   1.93 Saltwater   10
27   1.75 Saltwater   10
28   2.06 Saltwater   10
29   3.10 Saltwater    1
30   3.24 Saltwater    1
31   3.98 Saltwater    1
32   3.24 Saltwater    1
33   9.96 Saltwater  0.1
34  10.01 Saltwater  0.1
35   9.36 Saltwater  0.1
36  10.40 Saltwater  0.1
```

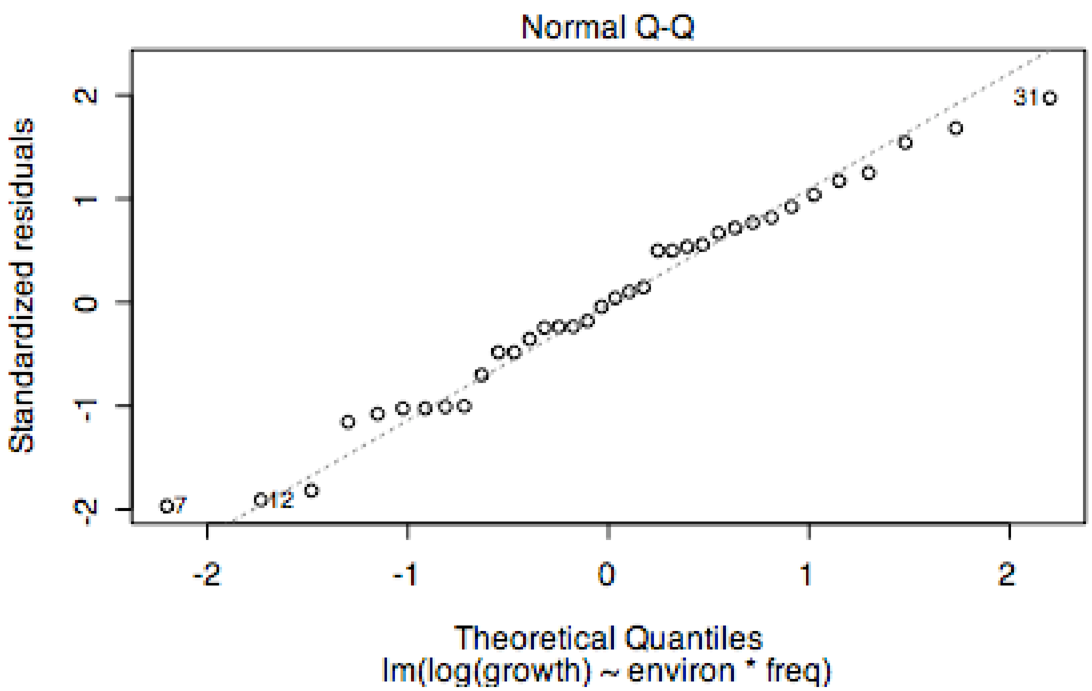
```
> anova(lm(growth~environ*freq, cracks))
Analysis of Variance Table

Response: growth
             Df  Sum Sq Mean Sq F value    Pr(>F)
environ       2  64.252  32.126  159.92 1.076e-15 ***
freq          2 209.893 104.946  522.40 < 2.2e-16 ***
environ:freq  4 101.966  25.491  126.89 < 2.2e-16 ***
Residuals    27   5.424   0.201
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> plot(lm(growth~environ*freq, cracks))
```






```
> anova(lm(log(growth)~environ*freq, cracks))
Analysis of Variance Table

Response: log(growth)
             Df Sum Sq Mean Sq F value    Pr(>F)    
environ       2 2.3576  1.1788 125.849 2.061e-14 ***
freq          2 7.5702  3.7851 404.095 < 2.2e-16 ***
environ:freq  4 3.5284  0.8821  94.172 1.885e-15 ***
Residuals    27 0.2529  0.0094                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> plot(lm(log(growth)~environ*freq, cracks))
```


Residuals vs Fitted
Residuals
0.2
0.1
0.0
-0.1
-0.2
31
12
1.0
1.5
2.0
Fitted values
lm(log(growth) ~ environ * freq)


Normal Q-Q
Standardized residuals
2
1
0
-1
-2
31
12
7
-2
-1
0
1
2
Theoretical Quantiles
lm(log(growth) ~ environ * freq)