MTH2006: Coursework 2

Ex 1a)

See appendix for code. We define our doses as factors and level them from “low” to “high”. We use arrange() function to order our table by doses.

Ex 1b)

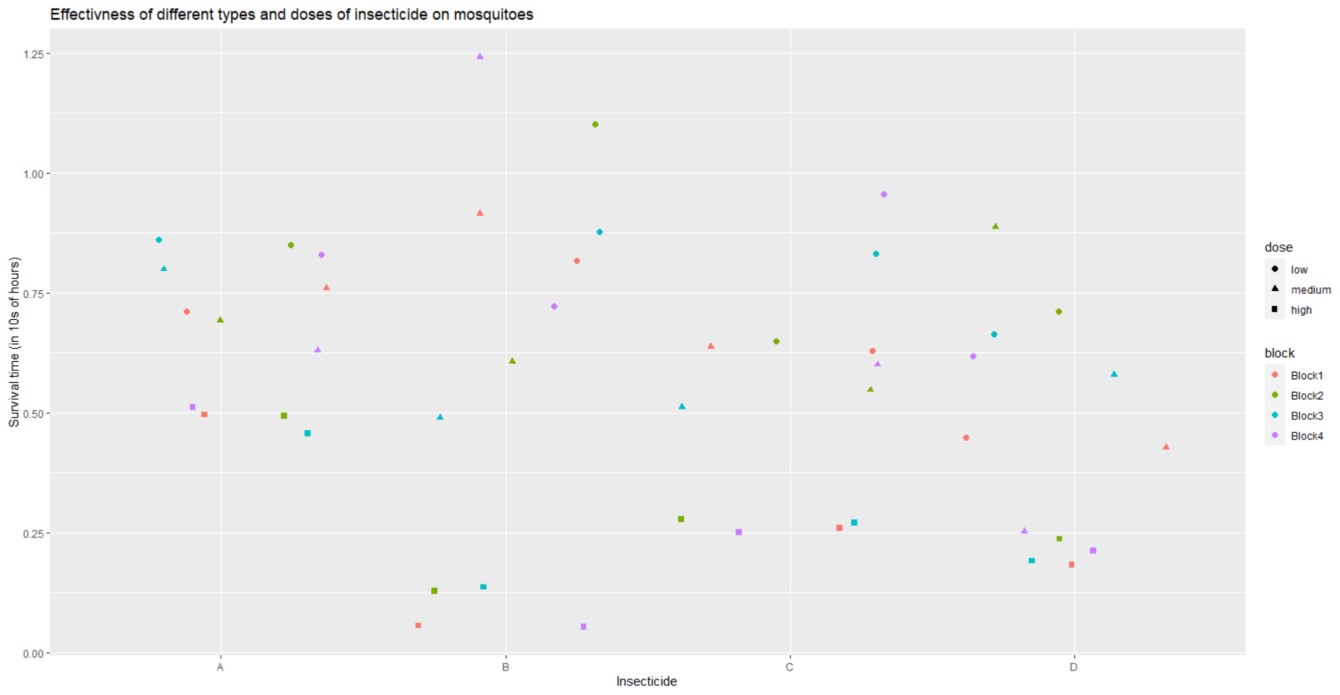

Figure 1 : effectiveness of different type and doses of insecticide on mosquitoes

From this plot, we see that doses appear to have a visible effect on survival time, with high doses having noticeably lower survival times. Insecticide A seems to have the smallest variation in survival time given the doses with B having the largest. C and D appear to have similar variation. Blocks do not seem to impact effectiveness of the insecticides.

Ex 1c)

In this study, 4 different types of insecticides were testes (A, B, C and D) using 3 different doses (low, medium, and high). The sample size was n=48, the experiment included 4 observations of each insecticide and dose from each block i.e there are 4 replicates.

Ex 1d)

Table 1: Summary of two-way analysis of variance table

	DF	Sum Sq	Mean Sq	F value	Pr (>F)
block	3	0.0304	0.0101	0.434	0.73029
insecticide	3	0.3126	0.1072	4.593	0.00856**
dose	2	2.2583	1.1291	48.369	1.55e-10***
insecticide:dose	6	0.3682	0.0614	2.628	0.03403*
Residuals	33	0.7704	0.0233		

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

We are using a treatment contrast with insecticide A with a low dose as our intercept. This approach seems appropriate as there is a natural reference point. The response variable is the survival time, the treatment factor is the insecticide, and the covariate is the dose. The ANOVA table indicates highly significant effects of the insecticide (p-value 0.008) and of the dose (p-value ~ 0) on the intercept. It also indicates a significant interaction between insecticide and dose (p-value 0.03) . The block appears to have no significant effect.

Table 2: Summary of linear model

```
Residuals:
   Min      1Q       Median    3Q      Max
-0.32500 -0.04875  0.00500  0.04312  0.42500

Coefficients:
                          Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)                0.81250   0.07457  10.896 5.98e-13 ***
insecticideB               0.06750   0.10546   0.640  0.52618
insecticideC              -0.04500   0.10546  -0.427  0.67213
insecticideD              -0.20250   0.10546  -1.920  0.06278 .
dosemedium                -0.09250   0.10546  -0.877  0.38623
dosehigh                  -0.32250   0.10546  -3.058  0.00419 **
insecticideB:dosemedium    0.02750   0.14914   0.184  0.85474
insecticideC:dosemedium    0.10000   0.14914  -0.671  0.50681
insecticideD:dosemedium    0.02000   0.14914   0.134  0.89407
insecticideB:dosehigh     -0.46250   0.14914  -3.101  0.00374 **
insecticideC:dosehigh     -0.18000   0.14914  -1.207  0.23533
insecticideD:dosehigh     -0.08250   0.14914  -0.553  0.58356
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1491 on 36 degrees of freedom
Multiple R-squared:  0.7864, Adjusted R-squared:  0.7211
F-statistic: 12.05 on 11 and 36 DF,  p-value: 4.96e-09
```

We fit the model with interaction between dose and insecticide. The model explains 79% of the variance in survival times. There is no significant distinction in the effect of different insecticides at low and medium doses, a high dose of insecticide however has a significant effect (reduce survival time by 0.32 compared to low dose). It also appears that a high dose of insecticide B significantly differs from the rest.

Ex 1e)

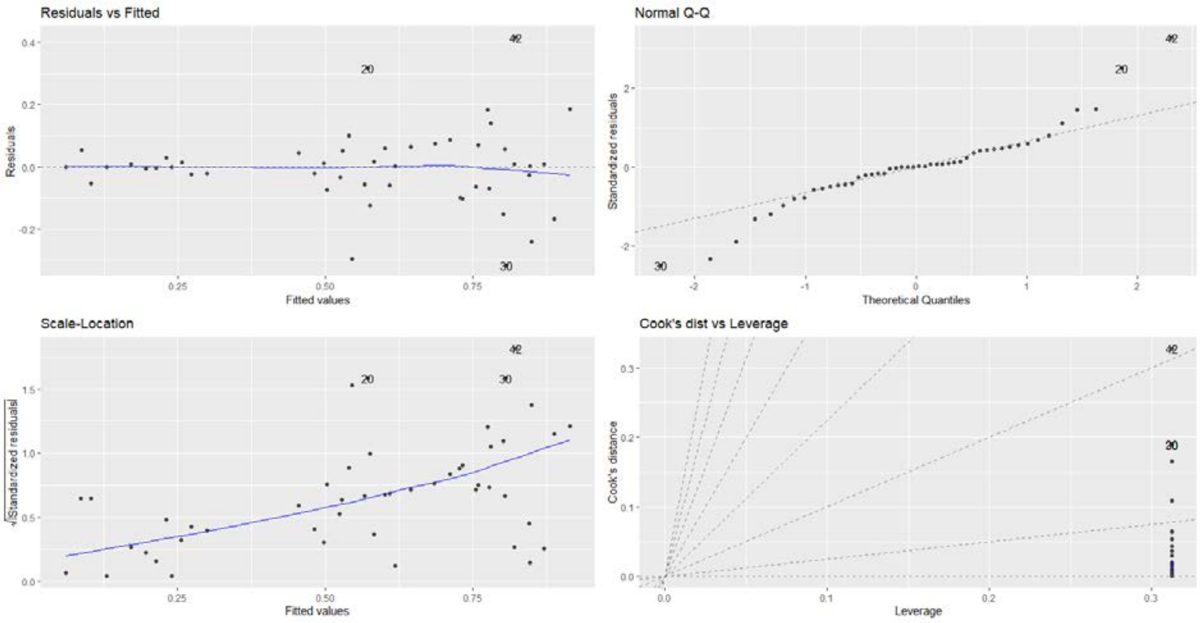

Figure 2: Diagnostic plot

Our residuals vs fitted plot seems to show the fit is good, the residuals should be approximately normally distributed with a mean of 0 and variance of 1, and shouldn't depend on the fitted values, which appears to be the case.

Ex 1f)

According to our model, the insecticide and dose which results in the lowest survival time is insecticide B with a high dose, an estimate of the survival time is 0.095 (in 10s of hours).

Ex 2a)

```
#Function to evaluate CDF
cdf <- function(y, alpha, beta){
  1-exp(-(beta*y)^alpha)
}

#Function to evaluate inverse of CDF
inv_cdf <- function(p, alpha, beta){
  (-log(1-p))^(1/alpha)/beta
}
```

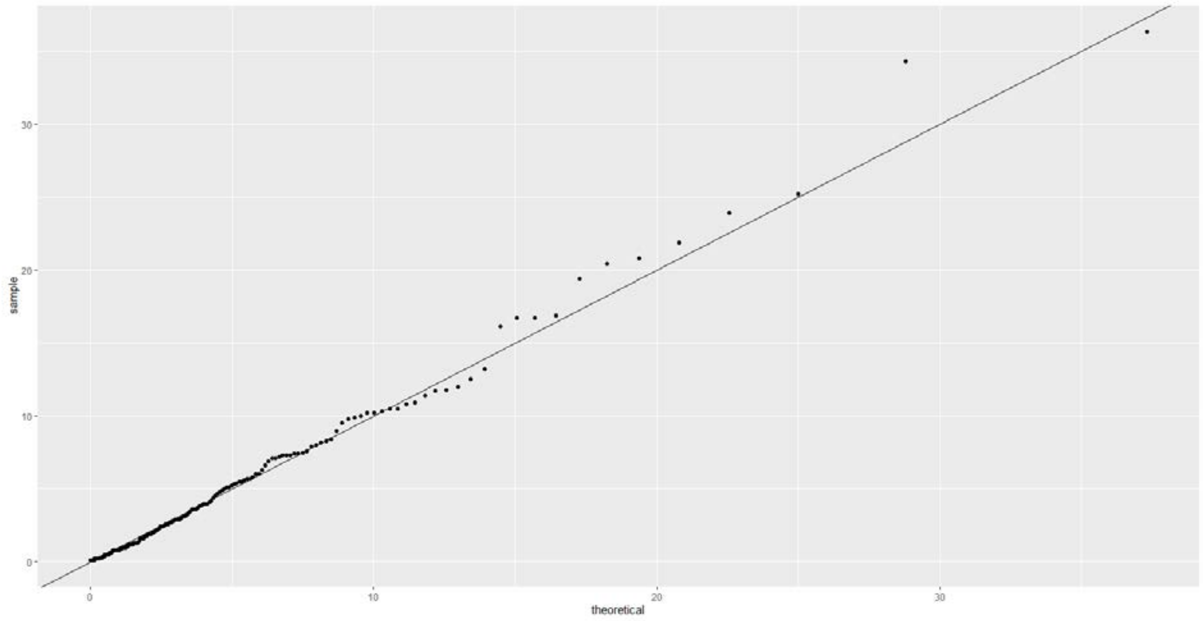

Figure 3: QQ-plot for $F_Y(y, \alpha, \beta)$ fit of Seavington rainfall

Figure 3 suggest our function is an okay fit to model non-zero rainfalls of Seavingtion. We do see some systematic deviation from the zero-intercept slope line for larger values of rainfaill, our theoretical values are lower than those observed.

Ex 2b)

Table 3 : Observed and Expected bin values

	[0,5)	[5,10)	[10,15)	[15,20)	[20,25)	[25,30)	[30,35)	[35,40)	[40,Inf)
Observed	**145.0**	**36.00**	**14.00**	**5.000**	**4.000**	**1.00**	**1.0000**	**1.0000**	**0.0000**
Expected	**148.8**	**34.59**	**13.01**	**5.547**	**2.539**	**1.22**	**0.6075**	**0.3112**	**0.3583**

We set our hypothesis:

$$H_0\text{: sample is from } F_Y(y,\alpha,\beta) \text{ distribution}$$

$$H_1\text{: sample is NOT from } F_Y(y,\alpha,\beta) \text{ distribution}$$

We have J = 9 bins, with two estimated parameters hence v = 6. χ^2_{obs} = sum((O - E)^2 / E) = 3.300546 (0 is observed values per bin and E expected)
The p-value for our test is $\Pr(\chi^2_6 > \chi^2_{obs}) = 0.770$ which strongly suggests not to reject the null hypothesis. The findings from our qqplot and our Chi-Squared test support each other, and we can say that $F_Y(y,\alpha,\beta)$ appears to be an okay fit for our model.

Ex 2c)

```
#Kernel Density square-root function

kde_sqrt <- function(y, y_i, w, h){
  d <- outer(sqrt(y), sqrt(y_i), FUN="-")
  rowMeans(w(d/h))/(h*2*sqrt(y))
}
```

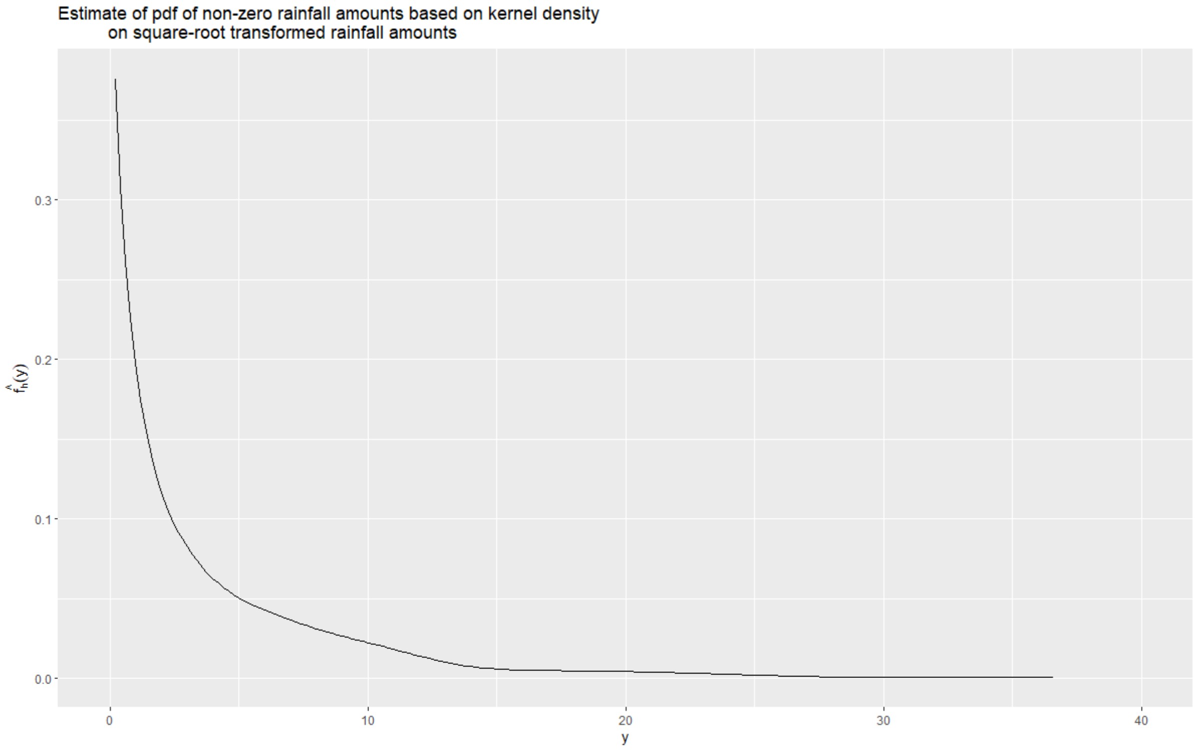

Figure 4: Estimate of pdf

Using kde_sqrt(c(0.5,1,2,5,10), non_zero$prcp, w_normal, h) we find :

```
0.22127007      0.16805552      0.11335383      0.05205675
```

APENDIX :

Code from Ex1 :

```
library(ggplot2)
library(dplyr)
#We read in our csv file
mosquito = read.csv("mosquito2021.csv")
#We make sure dose is considered as a factor from low to high
mosquito$dose = factor(mosquito$dose, levels = c("low", "medium", "high"))

#We arrange by dose
arrange(mosquito, dose)

#Making a plot to compare effectivness of insecticide
ggplot(data = mosquito, aes(x=insecticide, y=hours10, color = block, shape = dose))+
  geom_jitter(size=2) + ylab("Survival time (in 10s of hours)") + xlab("Insecticide") +
  ggtitle("Effectivness of different types and doses of insecticide on mosquitoes")

#Sample size
length(mosquito$X)

#AOV analysis exploration
summary(aov(hours10~dose, data = mosquito))
summary(aov(hours10~insecticide, data = mosquito))
summary(aov(hours10~block, data = mosquito))
summary(aov(hours10~dose+block, data = mosquito))
summary(aov(hours10~insecticide+block, data = mosquito))
summary(aov(hours10~dose+insecticide, data = mosquito))
summary(aov(hours10~dose+insecticide+block, data = mosquito))
summary(aov(hours10~insecticide*dose*block, data = mosquito))
summary(aov(hours10~dose+insecticide*block, data = mosquito))
summary(aov(hours10~dose*block+insecticide, data = mosquito))
summary(aov(hours10~dose*insecticide, data = mosquito))


#3Way test
summary(aov(hours10~dose*insecticide+block, data = mosquito))

#Final ANOVA choice
final_anova <- summary(aov(hours10~block+insecticide*dose, data = mosquito))
capture.output(final_anova, file = "anova results.doc")

#Linear model
lm_model <- summary(model <- lm(hours10~insecticide*dose, data = mosquito))
capture.output(lm_model, file = "lm_results.doc")
lm_model

#Fitted vs Residuals
```

```
library(ggfortify)
p <- autoplot(lm_model, which = c(1:3, 6)) # diagnostic plots
p[1] <- p[1] + ggtitle("Residuals vs Fitted")
print(p)
```

Code from Ex2 :

```
library(ggplot2)

#read our csv file
seavington = read.csv("seavington2020.csv")

#Non-zero rainfall amounts
non_zero = subset(seavington, prcp > 0)
non_zero_vec = non_zero[2]

#Function to evaluate CDF
cdf <- function(y, alpha, beta){
  1-exp(-(beta*y)^alpha)
}

#Function to evaluate inverse of CDF
inv_cdf <- function(p, alpha, beta){
  (-log(1-p))^(1/alpha)/beta
}

cdf_seavington = cdf(non_zero_vec, 0.775, 0.272)
inv_cdf_seavingtion = inv_cdf(cdf_seavington, 0.775, 0.272)

#Plotting QQ-plot
cdf_params = list(alpha = 0.775, beta = 0.272)
ggplot(non_zero, aes(sample = prcp)) +
  stat_qq(distribution = inv_cdf, dparams = cdf_params) + geom_abline()


#Pearsons Chi-Squared test
bins = c(0, 5, 10, 15, 20, 25, 30 , 35 ,40, Inf)
O <- table(cut(non_zero$prcp, bins, right = FALSE))
Phi <- cdf(bins, 0.775 , 0.272)
E <- length(non_zero$prcp) * diff(Phi)
J <- length(bins) - 1
contingency <- rbind(Observed = O, Expected = E)
print(contingency, digits = 4)
test_stat <- sum((O - E)^2 / E)
1 - pchisq(test_stat, J - 3)


#Kernel Density square-root function
```

```
kde_sqrt <- function(y, y_i, w, h){
  d <- outer(sqrt(y), sqrt(y_i), FUN="-")
  rowMeans(w(d/h))/(h*2*sqrt(y))
}


#Estimating Seavington rainfall
w_normal <- function(u) dnorm(u)
h <- 0.3
y_plot <- pretty(range(non_zero_vec) + h*c(-1,1), 200)
y_plot <- subset(y_plot, y_plot > 0)
f_hat <- kde_sqrt(y_plot, non_zero$prcp, w_normal, h)
ggplot() + geom_line(aes(x = y_plot, y = f_hat )) +
  labs(x = "y", y = expression(hat(f[h])(y))) + xlim(0, 40) +
  ggtitle("Estimate of pdf of non-zero rainfall amounts based on kernel density
        on square-root transformed rainfall amounts")

#Estimate f_yhat
kde_sqrt(c(0.5,1,2,5,10), non_zero$prcp, w_normal, h)
```