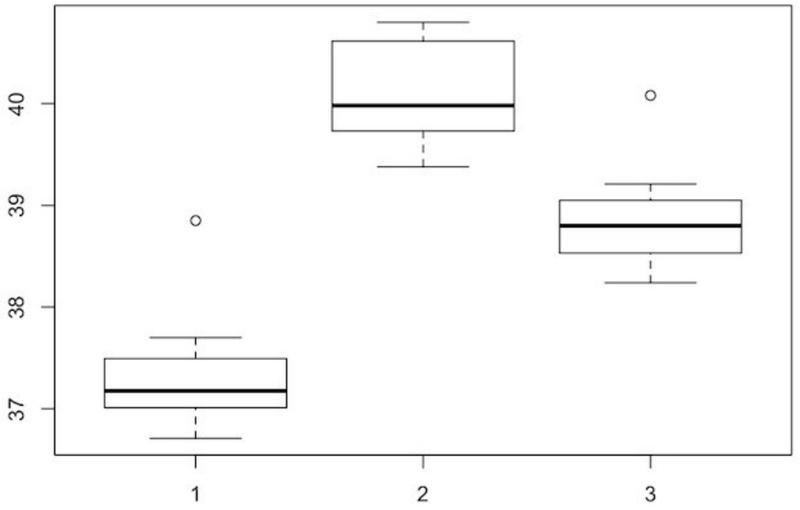
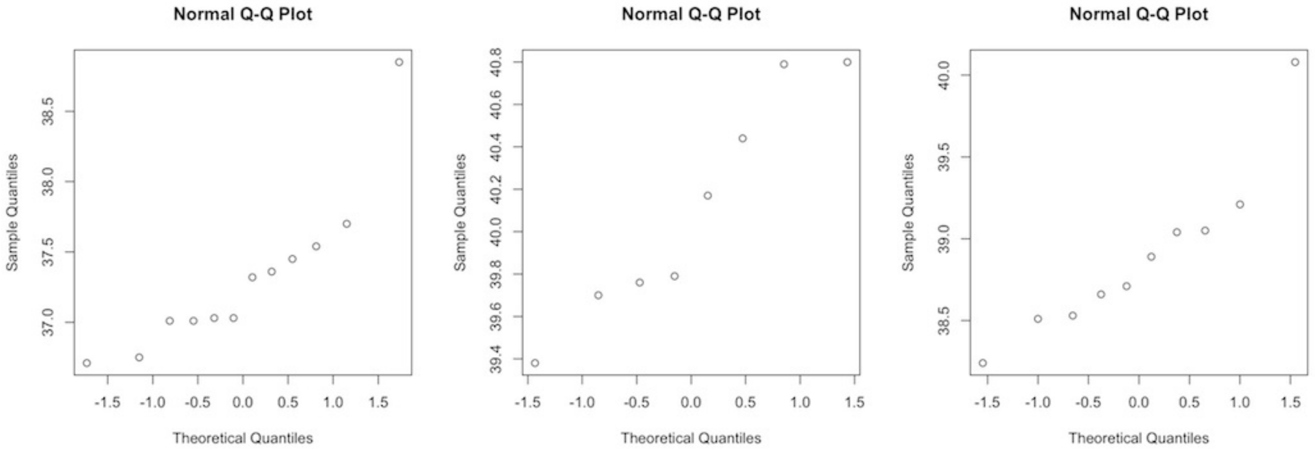
S520 Homework 10

Enrique Areyan
April 13, 2012

12.6.#A: For all of these questions, I wrote a function in R to calculate all of the data in the ANOVA table. The code is included as an appendix at the end of this homework.

1.





For the assumption of normality: the distribution that seems closer to normality is set B, although its median's location may question the symmetry (and hence) normality assumption. The other two distributions pretty much look symmetric, but have outliers that may question the normality assumption. I wouldn't jump to the conclusion that this data is normally distributed, but it may be sufficient to use ANOVA.

For the assumption of homoscedasticity: the boxplot suggests that data set A and C pretty much have the same variance, but data set B have a somewhat larger variance. I would conclude that the homoscedasticity assumption is plausible.

2.

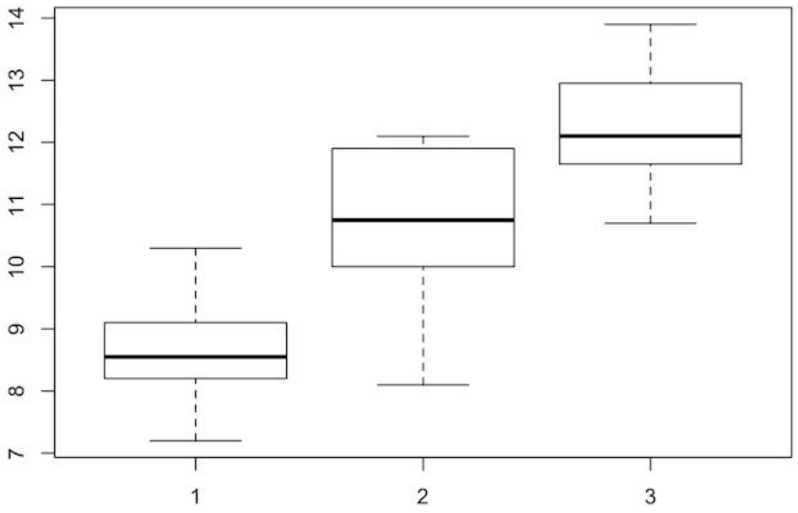
Source	SS	df	MS	F	**p**
Between	38.800883	2	19.4004413	66.02105	4.008649e-11
Within	7.934014	27	0.2938524		
Total	46.734897	29			

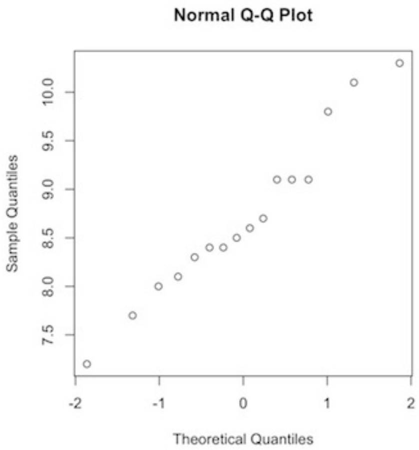
$$\mathbf{p} = 4.008649e - 11 < \alpha = 0.05 \Longrightarrow \text{ reject } H_0$$

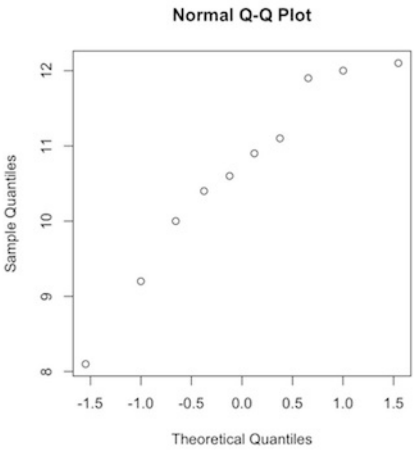
Somewhat similar to what was concluded in case study 12.5, "altought the assumptions of normality ... [is] suspect, the significance probability is so small that we feel comfortable rejecting the fundamental null hypothesis of equal population means".

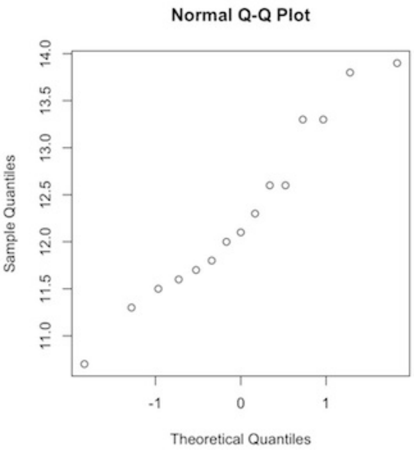
12.6.#B:

1.









For the assumption of normality: data set SS and SC seem to be as normal as it can get, while data set ST (the second box plot) is skew to the left. Thus, I would conclude with confidence that for the first and last data set the normality assumption hold, but I wouldn't be so sure about the second data set. For the assumption of homoscedasticity: the equal variance assumption is suspect for all three data set, but it may seem more plausible between data set SS (first) and SC (third).

2.

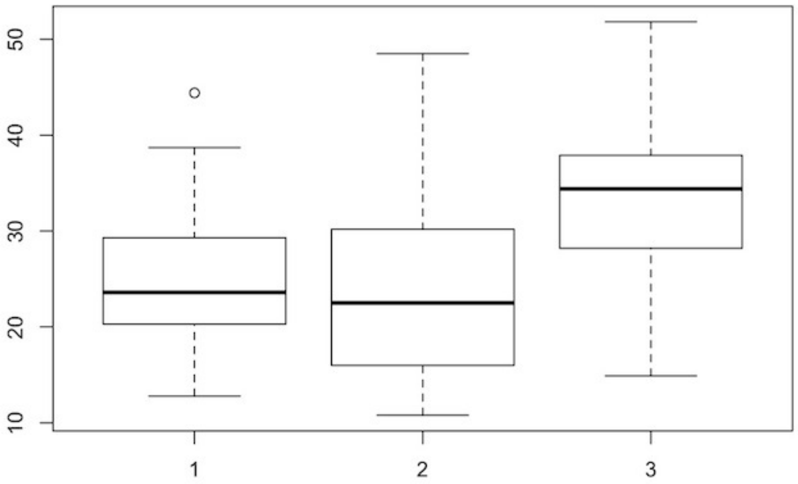
Source	SS	df	MS	F	**p**
Between	99.8893	2	49.9446524	49.99926	2.281786e-11
Within	37.9585	38	0.9989079		
Total	137.8478	40			

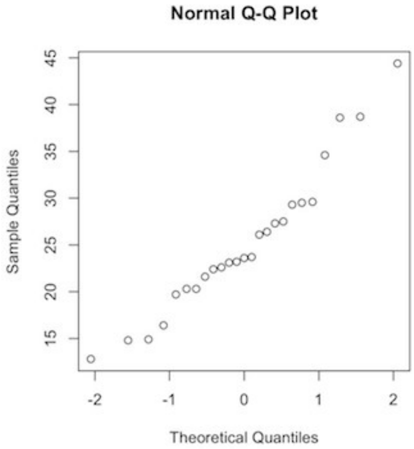
$$\mathbf{p} = 2.281786e - 11 < \alpha = 0.05 \Longrightarrow \text{ reject } H_0$$

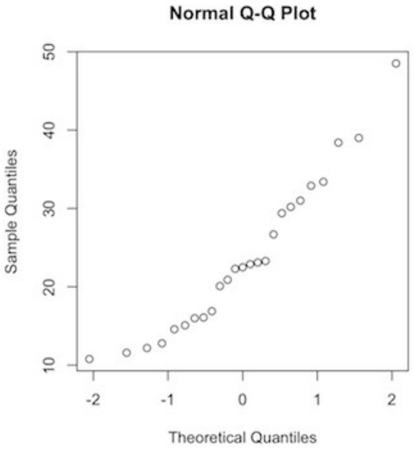
Somewhat similar to what was concluded in case study 12.5, "altought the assumptions of ... equal population variance [is] suspect, the significance probability is so small that we feel comfortable rejecting the fundamental null hypothesis of equal population means".

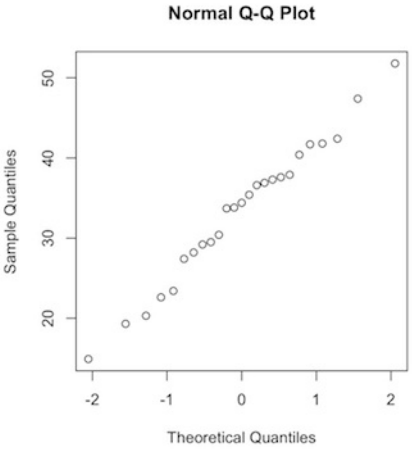
12.6.#E:

1.









For the assumption of normality: data set RS and NS seem to be pretty close to being normal distributed. Data set SS is skew to the right, and possibly fail to be symmetric and thus, normal distributed.
For the assumption of homoscedasticity: again, data set RS and NS seem to have a somewhat similar variance. Just looking at the box in the boxplot, there is not a lot of evidence against the homoscedasticity for all three distribution, but the second one may seem to have a slighter larger variation than the other two.

2. (i). Do the two selected lines (RS and SS) differ in fecundity from the nonselected line(NS)? Let μ_1 be the mean for RS, μ_2 for SS and μ_3 for NS. Then,

$$H_0 : \frac{\mu_1 + \mu_2}{2} = \mu_3 \text{ vs. } H_1 : \frac{\mu_1 + \mu_2}{2} \neq \mu_3 \Longrightarrow \theta_1 = 0 \text{ vs. } \theta_1 \neq 0$$

where $\theta_1 = \mu_1 + \mu_2 - 2\mu_3$, i.e., $c = (1, 1, -2)$

(ii). Do the line selected for resistance (RS) differ in fecundity from the line selected for susceptibility (SS)?

$$H_0 : \mu_1 = \mu_2 \text{ vs. } H_1 : \mu_1 \neq \mu_2 \Longrightarrow \theta_2 = 0 \text{ vs. } \theta_2 \neq 0$$

where $\theta_2 = \mu_1 - \mu_2 - 0\mu_3$, i.e., $c = (1, -1, 0)$

θ_1 and θ_2 are orthogonal, i.e. $1 \cdot 1 + 1 \cdot -1 + (-2) \cdot 0 = 1 - 1 = 0$ (recall that $n_1 = n_2 = n_3 = 25$)

To maintain a family rate of Type I error equal to 5%, we must maintain the following relation:

$$0.05 = 1-(1-\alpha)^2 \iff \alpha = 1-\sqrt{0.95} \approx 0.0253$$

So, we must perform each test at $\alpha = 0.0253$ in order to avoid a family error of more that 5%.

3. The ANOVA table is:

Source	SS	df	MS	F	**p**
Between	1362.211	2	681.10573	8.665739	0.000424431
Within	5659.022	72	78.59753		
Total	7021.234	74			

We need to calculate only SS_{θ_1}, and then use the fact that $SS_B = SS_{\theta_1} + SS_{\theta_2}$ to obtain SS_{θ_2}. Recall that in this case: $\theta_1 = (1, 1, -2)$.

$$SS_{\theta_1} = \frac{([1 \cdot 25.256 + 1 \cdot 23.628 + (-2) \cdot 33.372])^2}{1^2/25 + 1^2/25 + (-2)^2/25} = \frac{318.9796}{0.24} = 1329.082$$

The expanded ANOVA table is:

Source	SS	df	MS	F	**p**
Between	1362.211	2	681.10573	8.665739	0.000424431
θ_1	1329.082	1	1329.082	16.90997	0.000102737
θ_2	33.129	1	33.129	0.4215018	0.5182544
Within	5659.022	72	78.59753		
Total	7021.234	74			

The fundamental null hypothesis, i.e., $H_0 : \mu_1 = \mu_2 = \mu_3$ results in $\mathbf{p} = 0.000424431 < 0.05 = \alpha$, and thus, get rejected. The first contrast θ_1, tested at $\alpha = 0.0253$ also gets rejected, while the second contrast θ_2 (at the same α) does not get rejected.

In conclusion, the evidence in the data supports the hypothesis that the two selected lines (RS and SS) differ in fecundity from the nonselected line(NS) but, the line selected for resistance (RS) does not differ in fecundity from the line selected for susceptibility (SS).

```
#
#  This function constructs the ANOVA table for an arbitrary data set. data contain either
# the whole data set or a summary of the data. A summary of the data contains,
# for each data set, the number of observations, mean, and variance in that order (see function anova.parse.d
# For instance, if data is given as a summary, the following vector works for example 12.2
# c(25,9.783685,29.89214,20,10.908170,18.75800,20,15.002820,51.41654,65). The last position
# is the total number of observations, i.e., n1+n2+...+nk.
# If the data is not summary, then data contains all observations, and you need to provide the
# function with k, the number of ways to partition the data and vec, which contains k-1 pointers
# to split the data in the correct positions. You can specify whether the data is given as a summary or not u
# third parameter summary.
#
anova.table <- function(data,k,vec,summary = FALSE){
#Get the data
if(!summary){
parts <- anova.parse.data(data,k,vec)
}else{
parts <- data
}
totaln <- parts[length(parts)]
firstsum  <-  0
secondsum  <-0
h <- k-1
#Calculate observed value of ssb
```

```
for(i in 0:h){
firstsum  <- firstsum  + (parts[(i*3)+1] * parts[((i+1)*3)-1] * parts[((i+1)*3)-1])
secondsum <- secondsum + (parts[(i*3)+1] * parts[((i+1)*3)-1])
}
secondsum <- (secondsum * secondsum) / totaln
ssb <- firstsum - secondsum
#Calculate observed value of ssw
ssw <- 0
for(i in 0:h){
ssw <- ssw + ((parts[(i*3)+1] - 1) * parts[((i+1)*3)])
}
#Construct ANOVA table
anovaresults <- matrix(-1,nrow=3,ncol=5)
dimnames(anovaresults)[[1]] <- c("Between","Within","Total")
dimnames(anovaresults)[[2]] <- c("SS","df","MS","F","p")
#Source
anovaresults[1,1] <- ssb
anovaresults[2,1] <- ssw
anovaresults[3,1] <- ssb + ssw
#Degrees of freedom
anovaresults[1,2] <- k-1
anovaresults[2,2] <- totaln-k
anovaresults[3,2] <- totaln-1
#Mean Squares
anovaresults[1,3] <- ssb / anovaresults[1,2]
anovaresults[2,3] <- ssw / anovaresults[2,2]
#Test Statistic F
anovaresults[1,4] <- anovaresults[1,3]/anovaresults[2,3]
#Significance Probability
anovaresults[1,5] <- 1-pf(anovaresults[1,4],anovaresults[1,2],anovaresults[2,2])
#Return Table
return(anovaresults)
}
#
#  This function parses a vector to be used by anova.table
#    for an arbitrary data set. data contains all observations, k contains
# the number of partitions of the data, and vec contains k-1 pointer
# to partition the data correctly. This function returns, for each data set
# the number of observations, the mean and the variance. The last position is always
# the total number of observations, i.e., n1+n2+...+nk.
# For example, if we call anova.parse.data(c(1,2,3,4),2,c(2)), we get
# [1] 2.0 1.5 0.5 2.0 3.5 0.5 4.0.
#
anova.parse.data <- function(data, k, vec){
if(k != (length(vec)+1)){
stop("The parameter vec should be equal in length to k-1")
}else{
j <- 0
i <- 1
h <- k - 1
for(l in 1:h){
j <- j+1
i1<- j
j <- vec[i]
i<- i +1
i2<- j
#print(paste(i1,i2,sep=","))
assign(paste("part",l,sep=""),data[i1:i2])
assign(paste("n",l,sep=""),length(data[i1:i2]))
assign(paste("mean",l,sep=""),mean(data[i1:i2]))
assign(paste("var",l,sep=""),var(data[i1:i2]))
}
j <- j+1
l <- l+1
#print(paste(j,length(data),sep=","))
assign(paste("part",l,sep=""),data[j:length(data)])
assign(paste("n",l,sep=""),length(data[j:length(data)]))
assign(paste("mean",l,sep=""),mean(data[j:length(data)]))
assign(paste("var",l,sep=""),var(data[j:length(data)]))

valuesreturn <- c()
totaln <- 0
for(l in 1:k){
valuesreturn <- c(valuesreturn,get(paste("n",l,sep="")),get(paste("mean",l,sep="")),get(paste("var",l,sep="")
totaln <- totaln + get(paste("n",l,sep=""))
}
ret <- c(valuesreturn,totaln)
}
return(ret)
}
```