HW9

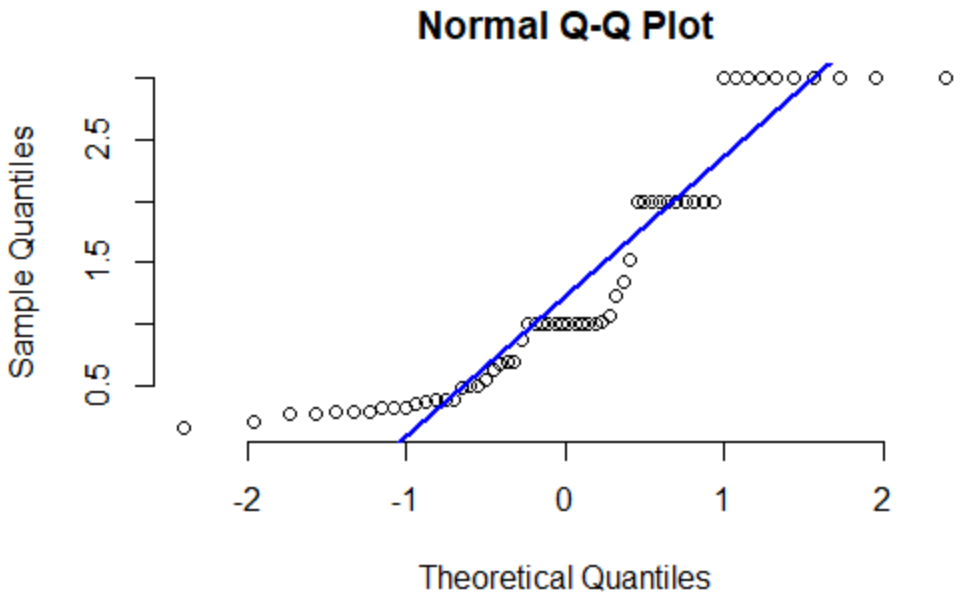
Code

Problem 9A

The concentrations (in nanograms per milliliter) of plasma epinephrine were measured for 10 dogs under isofluorane, halothane, and cyclopropane anesthesia; the measurements are given in the following table (Perry et al. 1974). Is there a difference in treatment effects? Use a parametric and a nonparametric analysis.

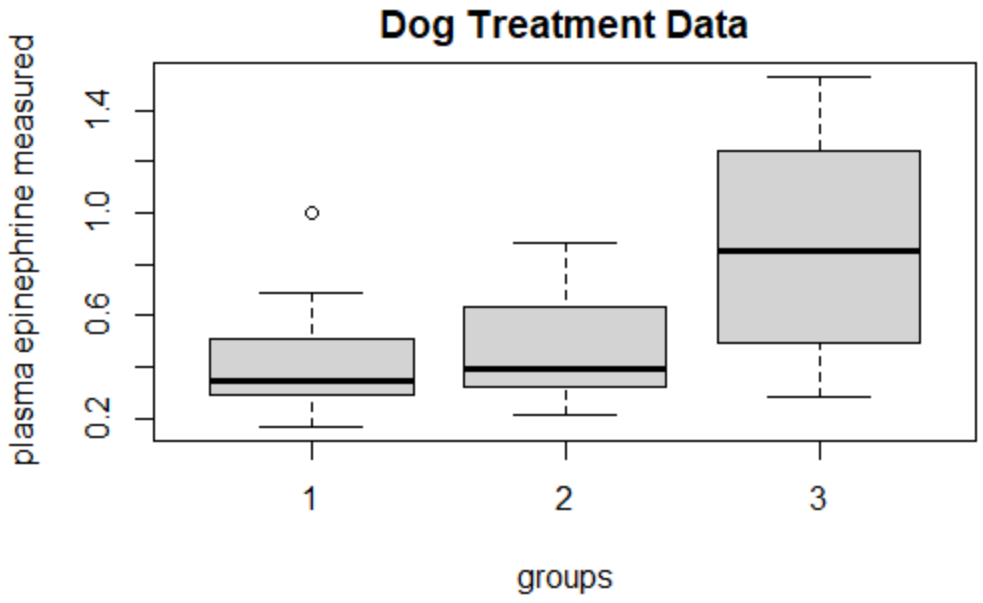
Hide

```
dogs <- read.table("hw9/dogs.txt", sep=" ")

dogs <- t(dogs)
dogs_row <- c(dogs[,1], dogs[,2], dogs[,3])
groups <- c(rep(1,10),rep(2,10),rep(3,10))
dogs <- cbind(dogs_row, groups)

qqnorm(dogs, pch=1, frame=FALSE)
qqline(dogs, col="blue",lwd=2)
```



Hide

```
boxplot(dogs_row~groups, data=dogs, main="Dog Treatment Data", ylab="plasma epinephrine measure
d")
```



The boxplots indicate that the mean and variances look different. Therefore, I will assume unequal variances.

Hide

```
oneway.test(dogs_row~groups, data=dogs)
kruskal.test(dogs_row~groups, data=dogs)
```

Given similar p-values of 0.05568 and 0.05948, it shows that the parametric and nonparametric results are the same. Because of this and the low p-values I would say taht there IS a difference in the treatment effects.

Problem 9B

Three species of mice were tested for “aggressiveness.” The species were A/J, C57, and F2 (a cross of the first two species). A mouse was placed in a 1-m2 box, which was marked off into 49 equal squares. The mouse was let go on the center square, and the number of squares traversed in a 5-min periodwas counted. Analyze the file C57, AJ, F2, using the Bonferroni method, to determine if there is a significant difference among species.

Hide

```
aj <- read.table("hw9/aj.txt", sep=" ")
c57 <- read.table("hw9/c57.txt", sep=" ")
f2 <- read.table("hw9/f2.txt", sep=" ")

boxplot(aj, data=aj)
boxplot(c57, data=c57)
boxplot(f2, data=f2)
```

The different boxplots indicate that there are unequal mean and variances. Because of this we are going to conduct CIs for each dataset and then compare them.

Hide

```
alpha=0.05

mu <- mean(aj$V1)
n <- length(aj$V1)
sd <- sd(aj$V1)
se <- sd/sqrt(n)
df =  n-1
t_score = qt(p=alpha/2, df=df,lower.tail=F)
error <- t_score*se

lower <- mu-error
upper <- mu+error
print(c(lower,upper))
```

For AJ, lower bound is 44.11812, upper bound is 64.59886.

Hide

```
alpha=0.05

mu <- mean(c57$V1)
n <- length(c57$V1)
sd <- sd(c57$V1)
se <- sd/sqrt(n)
df =  n-1
t_score = qt(p=alpha/2, df=df,lower.tail=F)
error <- t_score*se

lower <- mu-error
upper <- mu+error
print(c(lower,upper))
```

For C57, lower bound is 212.9886, upper bound is 250.2190.

Hide

```
alpha=0.05

mu <- mean(f2$V1)
n <- length(f2$V1)
sd <- sd(f2$V1)
se <- sd/sqrt(n)
df =  n-1
t_score = qt(p=alpha/2, df=df,lower.tail=F)
error <- t_score*se

lower <- mu-error
upper <- mu+error
print(c(lower,upper))
```

For C57, lower bound is 102.5120, upper bound is 157.9166.

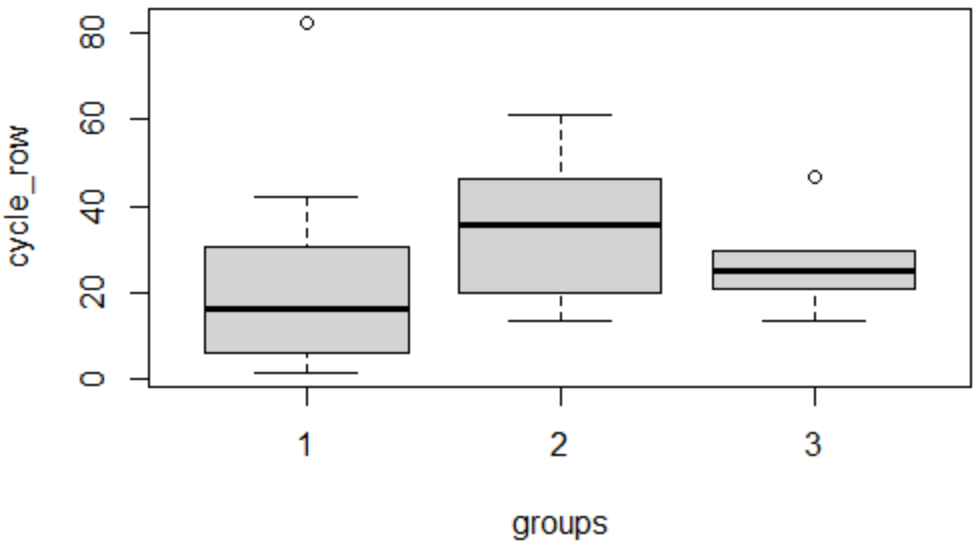
Given the fact that in the 95% confidence intervals for the three species that none of the bounds overlap at all, it is safe to assume that there IS a significant difference among species.

Problem 9C

Samples of each of three types of stopwatches were tested. The following table gives thousands of cycles (on-off-restart) survived until some part of the mechanism failed (Natrella 1963). Test whether there is a significant difference among the types, and if there is, determine which types are significantly different. Use both a parametric and a nonparametric technique.

Hide

```
Type_1 <- c(1.7, 1.9, 6.1, 12.5, 16.5, 25.1, 30.5, 42.1, 82.5)
Type_2 <- c(13.6, 19.8, 25.2, 46.2, 46.2, 61.1)
Type_3 <- c(13.4, 20.9, 25.1, 29.7, 46.9)
cycle_row <- c(Type_1, Type_2, Type_3)
groups <- c(rep(1,9),rep(2,6),rep(3,5))
cycles <- cbind(cycle_row, groups)


boxplot(cycle_row~groups, data=cycles)
```



The boxplots of the three types show that they might have equal means, and most likely unequal variances. The sample sizes are quite small.

Hide

```
oneway.test(cycle_row~groups, data=cycles)
```

```
    One-way analysis of means (not assuming equal
    variances)

data:  cycle_row and groups
F = 0.5206, num df = 2.000, denom df = 11.013, p-value
= 0.6081
```

Hide

```
kruskal.test(cycle_row~groups, data=cycles)
```

```
    Kruskal-Wallis rank sum test

data:  cycle_row by groups
Kruskal-Wallis chi-squared = 2.1547, df = 2, p-value =
0.3405
```

Given the p-values of 0.6081 and 0.3405, it would be difficult to reject the hypothesis even with such small sample sizes. So, we fail to reject hypothesis of the “same distribution”.

Problem 9D

a. Show that $R(X,Y) = R(Y,X) = R(X^*,Y^*)$. This means that the correleation between two variables doesn’t depend on the order of the variables nor on the units in which the variables where measured.

First equality is true because of multiplication rules. Since $E(X^*) = E(Y^*) = 0$ and $SD(X^*) = SD(Y^*) = 1$. Because of this, the definition says that $R(X^*,Y^*) = E(X^*Y^*) = R(X,Y)$.

b. Show that $-1 \leq R(X,Y) \leq 1$.

Given the hint that $(X^* + Y^*)^2$ and $(X^* - Y^*)^2$ are non-negative, we can say that the expectations:

$E[(X^* + Y^*)^2] \geq 0 \Rightarrow E[X^{*2} + 2X^*Y^* + Y^{*2}] = 1 + 2p + 1 \geq 0 \Rightarrow p \geq -1$
$E[(X^* - Y^*)^2] \geq 0 \Rightarrow E[X^{*2} - 2X^*Y^* + Y^{*2}] = 1 - 2p + 1 \geq 0 \Rightarrow p \leq 1$

Problem 9E

Some define r as: $r = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{x_i-\bar{x}}{s_x})(\frac{y_i-\bar{y}}{s_y})$ Show that this is equivalent to $r = \frac{1}{n}\sum_{i=1}^{n}(\frac{x_i-\bar{x}}{\sigma_x})(\frac{y_i-\bar{y}}{\sigma_y})$

Since $\sigma^2 = \frac{n-1}{n}s^2$,

$r = \frac{1}{n}\sum(\frac{x_i-\bar{x}}{s_x\sqrt{\frac{n-1}{n}}})(\frac{y_i-\bar{y}}{s_y\sqrt{\frac{n-1}{n}}})$

$= \frac{1}{n}\frac{n}{n-1}\sum(\frac{x_i-\bar{x}}{s_x})(\frac{y_i-\bar{y}}{s_y})$

$\frac{1}{n-1}\sum(\frac{x_i-\bar{x}}{s_x})(\frac{y_i-\bar{y}}{s_y})$

Showing that the two expressions are equivalent.

Problem 9F

A more comprehensible form of the equation of the regression line for estimating y based on x. In class we derived expressions for the slope and the intercept of the regression line.

a. Show that the slope of the regression line is $\hat{b} = r\frac{\sigma_y}{\sigma_x} = r\frac{s_y}{s_x}$

$\hat{b} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$
$= \frac{Cov(x,y)}{\sigma_x^2}$
$= \frac{r\sigma_x\sigma_y}{\sigma_x^2}$
$= r\frac{\sigma_y}{\sigma_x}$
$= r\frac{s_y}{s_x}$

b. Show that the intercept of the regression line is $\hat{a} = \bar{y} - \hat{b}\bar{x}$, and that the regression line passes through the point of averages $(\bar{x}, \bar{y})$.

Equation of the regression line is $\hat{y} = \bar{y} - \hat{b}\bar{x} + \hat{b}x$. After plugging in the $\bar{x}$ for x, then $\hat{y} = \bar{y}$, therefore the line goes through $(\bar{x}, \bar{y})$.

c. Show that the equation of the regression line can be written as $(\frac{\hat{y}-\bar{y}}{\sigma_y}) = r(\frac{\hat{x}-\bar{x}}{\sigma_x})$

Given the equation line before, $\hat{y} - \bar{y} = \hat{b}(x_i - \bar{x}) \Rightarrow \hat{y} - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x_i - \bar{x}) \Rightarrow (\frac{\hat{y}-\bar{y}}{\sigma_y}) = r(\frac{x-\bar{x}}{\sigma_x})$

d. The “regression effect”

If the student is 1.5 SDs above avg on the midterm, the SU of the midterm score is 1.5 units. Therefore, the predicted final score is 1.5r. Since the midterm and final score are positively correlated, but not perfectly linear, we expects the finald score to be 1.5SDs or less on average. Same if it was the other way around, -1.5r in SU would be 1.5SDs or better below the mean.

Problem 9G

The fitted values. The estimated values ˆy are often called the f itted values because they are obtained by fitting the regression model to the data. Use the fact that ˆy is a linear function of x to show that:

a. $\bar{\hat{y}}$. That is, the average of the fitted values equals the average of the original values.

$\frac{1}{n}\sum\hat{y}_i$
$= \frac{1}{n}\sum(\hat{a} + \hat{b}x_i)$
$= \hat{a} + \hat{b}\bar{x}$
$= \bar{y} - \hat{b}\bar{x} + \hat{x}\bar{x}$
$= \bar{y}$

b. $\sigma^2_{\hat{y}} = r^2\sigma^2_y$. Check that this gives sensible answers when r = 0 and when r = 1.

Variance, by definition, is $\sigma^2_{\hat{y}} = \frac{1}{n}\sum(\hat{y}_i - \bar{y})^2$

Therefore, $\sigma^2_{\hat{y}}$

$$= \frac{1}{n}\sum(\hat{a} + \hat{b}x_i - \bar{y})^2$$
$$= \frac{1}{n}\sum(\bar{y} - \hat{b}\bar{x} + \hat{b}x_i - \bar{y})^2$$
$$= \frac{1}{n}\sum(\hat{b}(x_i - \bar{x}))^2$$
$$= \hat{b}^2\sigma_x^2$$
$$= r^2\sigma_y^2$$

When regression is flat, $\hat{y} = \bar{y}$, so their var is zero since the fitted values are the same. Similarly, when regressions is 1, the points sit on a regressions line–all fitted values are the same and the variances are the same.

Problem 9H

The residuals. For each $i = 1, 2, 3, \ldots, n$, define $\hat{e}_i = \hat{y}_i - y_i$ to be the ith residual, that is, the error in the regression estimate of y_i.

a. Show that $\bar{\hat{e}} = 0$.

$\bar{\hat{e}} = \frac{1}{n}\sum(\hat{y}_i - y_i)$, when all the residuals are summed, then $\bar{\hat{y}} = \bar{y}$, therefore $\frac{1}{n}\sum(\hat{y}_i - y_i) = 0$

b. Show that $\sigma^2_{\hat{e}} = (1 - r^2)\sigma_y^2$.

Since $\bar{\hat{e}} = 0$,

$\sigma^2_{\hat{e}} = \frac{1}{n}\sum\hat{e}_i^2 = \frac{1}{n}\sum(\bar{y} - \hat{b}\bar{x} + \hat{b}x_i - y_i)^2$

Can simplify the expression $\hat{b}^2\sigma_x^2 - 2\hat{b}r\sigma_x\sigma_y + \sigma_y^2 = \sigma_y^2(r^2 - 2r^2 + 1) = \sigma_y^2(1 - r^2)$

When r = 0 or 1, the residuals cause the variances to either be 0 or the mean squared residual is the variance.

Problem 9I

Decomposing the sum of squares.

a. Show that $\sigma_y^2 = \sigma^2_{\hat{e}} + \sigma^2_{\hat{y}}$

$\sigma^2_{\hat{e}} + \sigma^2_{\hat{y}} = \sigma_y^2(1 - r^2) + r^2\sigma_y^2 = \sigma_y^2$

b. Show that $\sum(y_i - \bar{y})^2 = \sum(y_i - \hat{y}_i)^2 = \sum(\hat{y}_i - \bar{y})^2$.

This is the exact same as in part a) except but that each variable is multiplied by $\frac{1}{n}$,

$\frac{1}{n} * [\sigma^2_{\hat{e}} + \sigma^2_{\hat{y}} = \sigma_y^2(1 - r^2) + r^2\sigma_y^2 = \sigma_y^2]$

Problem 9J

The residuals and x.

a. Show that $\sum(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$

$(y_i - \bar{y}^2)$
$= [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2$
$= (y_i - \hat{y}_i) + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + (\hat{y} - \bar{y})^2$

Cross-product term is 0, supported by what we saw in 9b.

b. Show that the residuals and x are uncorrelated. This explains why the residual plot (a plot of the residuals versus x) should show no trend upwards or downwards.

Uncorrelated means corr = 0. By this,

$Cov(x, \hat{e}) = \frac{1}{\hat{b}} \frac{1}{n} \sum(\hat{y}_i - \bar{y})(\hat{y}_i - y_i) = 0.$