Question 1.

(a) Describe the study design being used to address the question of interest. What are the treatment groups and their corresponding sample sizes?

(b) What are the experimental and measurement units for this study?

(c) Write down an appropriate statistical model for these data and describe its assumptions in the context of this study. What is the hypothesis that the investigators are interested in testing?

Solution 1. *For (a):*

This is a completely randomized (with no block) experiment. The objective is to understand the following:

(1) If the acid rain harms the tree or not

(2) If the answer is 'Yes', does it depends on the pH value of the rain.

The treatment factors are five different pH values assigned to five groups of yellow birch seedlings (experimental units). So seedlings given the same pH treatment form a treatment group, and the sample sizes of all groups are the same, which is $240/5 = 48$.

For (b):

In this case, the experimental and measurement units are the same, both are the yellow birch seedlings.

For (c):

The statistical model would be a completely randomized design (I am using the book's notation here):

$$\begin{aligned} Y_{it} &= \mu + \tau_i + \epsilon_{it}, \\ \epsilon_{it} &\sim N(0, \sigma^2) \\ \epsilon'_{it}s &\text{ are i.i.d,} \\ i &= 1, \ldots, 5, t = 1, \ldots, 48 \end{aligned}$$

Here Y_{it} *is the total plant (dry) weight measured after 17 weeks(response).* μ *is the mean value of the dry weight without acid rain treatment.* τ_i *is the impact of the i-th ph value to the weight of the seedlings, and* ϵ_{it} *is the error variable (minor source of variation).* σ^2 *is an unknown constant representing the variance of the error.*

The hypothesis here is:

$$\begin{aligned} H_0 &: \tau_1 = \tau_2 = \ldots = \tau_5 s \\ H_a &: \tau_i \neq \tau_j \text{ for some } i \neq j \end{aligned}$$

H_0 *is translated as the impact of pH values on the weight of seedlings are the same.* H_1 *is translated as at least two different pH values do different harms to the seedlings.*

Question 2.

Solution 2. *The R code for this question is attached as HW1Q2.R. It is self contained with comments. In our solution to the questions, we will quote the line number of the code wherever needed to support our results.*

For (a):

We design a completely randomized experiment with two treatment factors (A and B). The experimental units are the 185 *male patients aged from* 50 *to* 70 *years old and the measurement units are their blood pressure after* 12 *weeks.*

So the design is a one way ANOVA model:

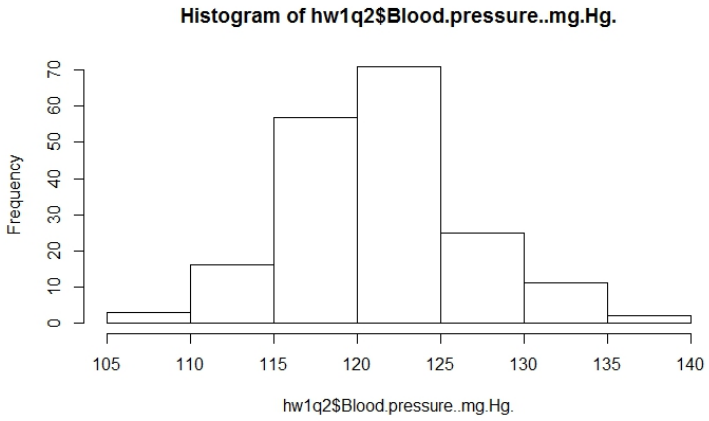
$$Y_{it} = \mu + \tau_i + \epsilon_{it}$$
$$\epsilon_{it} \sim N(0, \sigma^2)$$
$$\epsilon'_{it}s \text{ are } i.i.d$$

Here Y_{it} *is the blood pressure measured after* 12 *weeks(response),* μ *is the mean blood pressure without treatment,* τ_i *is the mean impact on blood pressure for the i-th treatment (*$i = 1$ *means treatment A,* $i = 2$ *means treatment B),* ϵ_{it} *is the error variable (covariates, etc).*

The number of levels of treament is $v = 2$*, so* $i = 1, 2$*. In each level, we have numbers of observations* $r_1 = 90, r_2 = 95$*(R code line* $1 - 19$*). The total number of observation is then* $n = r_1 + r_2 = 185$*.*

*In the model above, we assumed independence, normality and constant variance for our response. Independence is met since the blood pressure is measured among different patients. Normality should be supported by relatively large sample. Constant variance is met since we are focusing on a particular age group (*50 *to* 70 *years) and gender (male only), also human blood pressure tends to stay within the same range(providing it is wide enough)*

To check normality roughly, we sketched a histogram on the blood pressure (R code line $21 - 22$*)*

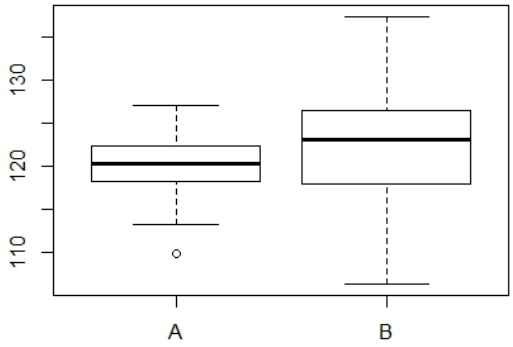


It does approximately look like fitting into the pattern of normal distribution. Our hypothesis is as following:

$$H_0 : \tau_1 = \tau_2$$
$$H_a : \tau_1 \neq \tau_2$$

H_0 is interpreted as treatment A and treatment B do not make a difference on patient blood pressure. H_a is the opposite statement as H_0.

From the result of ANOVA model, we know that under the null hypothesis, $\frac{MST}{MSE} \sim F_{v-1,n-v} = F_{1,183}$. First we run R code line $24 - 25$ to get a boxplot:



which shows support for rejecting the null.

Then we run R code line $27 - 32$ to get ANOVA table:

```
Analysis of Variance Table

Response: hw1q2a$Blood.pressure..mg.Hg.
                  Df Sum Sq Mean Sq F value
hw1q2a$Treatment   1  203.8 203.836  7.3943
Residuals        183 5044.7  27.567
                   Pr(>F)
hw1q2a$Treatment 0.007173 **
Residuals
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

with p value as 0.007173, we consider it small enough (smaller than 0.01) to reject the null hypothesis

So the conclusion is that we think there is a difference of effect on the patients' blood pressure between the two treatments.

Thus finished part (a)*.*

For part (b)*:*

To do this part we implemented two R packages: mosaic and ggplot2(R code line $3-4$*)*

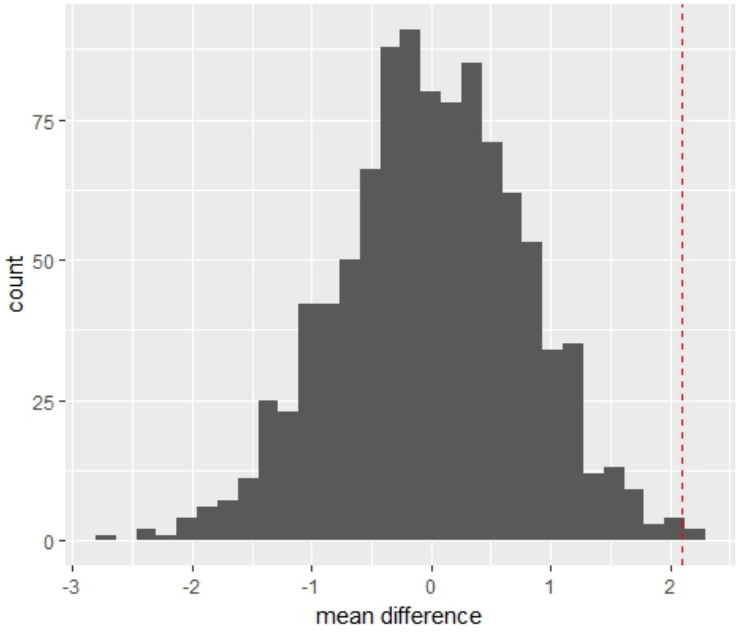
The hypothesis is the following:

$$H_0 : \mu_A = \mu_B$$
$$H_a : \mu_A \neq \mu B$$

μ_A, μ_B *separately represents the sample mean of the blood pressure measurement in each treatment.*

Please refer to R code line $34-60$ *for this part.*

It needs to be pointed out that there are $\frac{185!}{90!95!}$ *different permutations in the randomization process, but to save time we are only simulating with* 1000 *trials.However from the histogram plot we see that it is good enough to approximate the real sample distribution*



The p-value we get here is 0.003*(the sub area of the bell to the right of red line), which is good enough for us to reject* H_0*, so we still get the same conclusion as in* (a) *that there is a difference between treatment A and B.*

Question 3.

Solution 3. *Denote* y_{it} *as the observed response, where* $i = 1, 2, 3, 4$*. Here the number of treatment factors is* $v = 4$*. Also* $t = 1, 2, \ldots, r_i$*, where* r_i *is the number of observations in treatment* i*. So* $r_1 = 7, r_2 = 8, r_3 = 6, r_4 = 8$*, and the total number of observations across different treatment group is* $n = \sum_i r_i = 29$*.*

So For (a)*:*

The overall mean is:

$$\bar{y}_{..} = \frac{1}{n}\sum_i \sum_t y_{it}$$

Please see R code line $1-24$ *to read in the data as the form we need, and code line* $28-29$ *to get the overall mean* $\bar{y}_{..} = 3.718276$.

Please see R code line $31-32$ *to get the treatment mean:*

$$\bar{y}_{1.} = 3.745714$$
$$\bar{y}_{2.} = 3.580000$$
$$\bar{y}_{3.} = 3.598333$$
$$\bar{y}_{4.} = 3.922500$$

Finally the treatment effects are the treatment mean minus the overall mean. See R code line $34-35$*:*

$$\textit{Treatment1 effect}: 0.02743842$$
$$\textit{Treatment2 effect}: -0.13827586$$
$$\textit{Treatment3 effect}: -0.11994253$$
$$\textit{Treamtnet4 effect}: 0.20422414$$

For part (b):

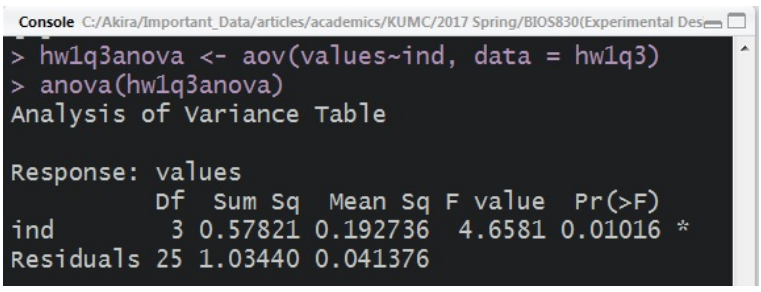
The general anova table assumes the following format:

Source	*DF*	*SS*	*MS*	*F*	*p*
Treatments	$v-1$	ssT	$msT = \frac{ssT}{v-1}$	$\frac{msT}{msE}$	
Error	$n-v$	ssE	$msE = \frac{ssE}{n-v}$		
Total	$n-1$	$sstot$			

In particular, we have:

$$ssE = \sum_i \sum_t y_{it}^2 - \sum_i r_i \bar{y}_{i.}^2$$
$$sstot = \sum_i \sum_t y_{it}^2 - n\bar{y}_{..}^2$$
$$ssT = \sum_i r_i \bar{y}_{i.}^2 - n\bar{y}_{..}^2$$

See R code line $37-39$*, we got:*



The ind row gives us the treament row for ANOVA table, and the Residuals row gives us Error row for ANOVA table. And we only need to add on column DF and SS to get the total row for ANOVA table. So the final ANOVA table is:

Source	*DF*	*SS*	*MS*	*F*	*p*
Treatments	3	0.57821	0.192736	4.6581	*0.01016*
Error	25	1.03440	0.041376		
Total	28	1.61261			

Since the p value is reasonably small (< 0.05), we conclude that the four diets do make a difference on the liver weight as a percentage of the body weight.

Question 4.

Solution 4. *For part (a), to show S_i^2 is an unbiased estimator of σ^2, only need to show that $E[S_i^2] = \sigma^2$. See the following work:*

$$\begin{aligned}
E[S_i^2] &= E\Big[\frac{1}{n_i - 1}\sum_j \Big(Y_{ij} - \bar{Y}_{i\cdot}^2\Big)^2\Big] \\
&= \frac{1}{n_i - 1}\sum_j E\Big[Y_{ij}^2 - 2Y_{ij}\cdot\bar{Y}_{i\cdot} + \bar{Y}_{i\cdot}^2\Big] \\
&= \frac{1}{n_i - 1}\Big\{\sum_j \underbrace{E[Y_{ij}^2]}_{=\sigma^2} - \sum_j \frac{2}{n_i}\Big(\sum_k E\Big[Y_{ij}Y_{ik}\Big]\Big) + n_i \underbrace{E\Big[\bar{Y}_{i\cdot}^2\Big]}_{=\sigma^2/n_i}\Big\}
\end{aligned}$$

Notice that due to the independence and normality assumption, we have:

$$\begin{aligned}
E\Big[Y_{ij}Y_{ik}\Big] &= E\Big[Y_{ij}\Big]\cdot E\Big[Y_{ik}\Big] = 0 \ \textit{if } j \neq k \\
E\Big[Y_{ij}Y_{ik}\Big] &= E\Big[Y_{ij}^2\Big] = \sigma^2 \ \textit{if } j = k
\end{aligned}$$

So continue from above we have:

$$\begin{aligned}
E[S_i^2] &= \frac{1}{n_i - 1}\Big\{\sum_j \sigma^2 - \sum_j \frac{2}{n_i}\sigma^2 + n_i\cdot\frac{\sigma^2}{n_i}\Big\} \\
&= \frac{1}{n_i - 1}\Big\{n_i\sigma^2 - n_i\cdot\frac{2}{n_i}\sigma^2 + \sigma^2\Big\} \\
&= \frac{1}{n_i - 1}\cdot(n_i - 1)\cdot\sigma^2 \\
&= \sigma^2
\end{aligned}$$

Thus finished the proof.

For part (b):

We can use the result from part (a):

$$\begin{aligned} E\Big[SSE_F\Big] &= E\Big[\sum_i (n_i - 1)S_i^2\Big] \\ &= \sum_i (n_i - 1)E\Big[S_i^2\Big] \\ &= \sum_i (n_i - 1)\sigma^2 \\ &= \sigma^2 \sum_i (n_i - 1) \\ &= (n - g)\sigma^2 \end{aligned}$$

Here $n = \sum_i n_i$ and we were given $i = 1, 2, \ldots, g$.

Question 5.

Solution 5. *For part (a):*

Notice that the error sum of square for the reduced model is:

$$SSE_0 = \sum_i \sum_j \Big(Y_{ij} - \hat{\mu}\Big)^2$$

So in order to get the least sum of square error, we take derivative on $\hat{\mu}$ and set it to 0:

$$\begin{aligned} \frac{\partial}{\partial \mu} SSE_0 = -\sum_i \sum_j 2\Big(Y_{ij} - \hat{\mu}\Big) = 0 \\ \implies \sum_i \sum_j \Big(Y_{ij} - \hat{\mu}\Big) = 0 \\ \implies \sum_i \sum_j Y_{ij} - \sum_i \sum_j \hat{\mu} = 0 \\ \implies \sum_i \sum_j Y_{ij} = n\hat{\mu} \\ \implies \hat{\mu} = \frac{1}{n} \sum_i \sum_j Y_{ij} = \bar{Y}_{\cdot\cdot} \end{aligned}$$

For part (b):

Follow the hint let's first compute $E\Big[SSE_0\Big]$:

$$\begin{aligned} E\Big[SSE_0\Big] &= E\Big[\sum_i \sum_j \Big(Y_{ij} - \bar{Y}_{\cdot\cdot}\Big)\Big]^2 \\ &= \sum_i \sum_j \Big[E\Big(Y_{ij}^2\Big) - 2E\Big(Y_{ij}\bar{Y}_{\cdot\cdot}\Big) + E\Big(\bar{Y}_{\cdot\cdot}^2\Big)\Big] \\ &= \sum_i \sum_j \Big[\sigma^2 - \frac{2}{n} E\Big[Y_{ij} \sum_k \sum_l Y_{kl}\Big] + \frac{\sigma^2}{n}\Big] \end{aligned}$$

Notice that due to the independence and nomrality of Y_{ij}, we have:

$$E\Big[Y_{ij}Y_{kl}\Big] = E[Y_{ij}]E[Y_{kl}] = 0 \text{ if } (i,j) \neq (k,l)$$
$$E\Big[Y_{ij}Y_{kl}\Big] = E[Y_{ij}^2] = \sigma^2 \text{ if } (i,j) = (k,l)$$

So continue from above, we have:

$$\begin{aligned} E\Big[SSE_0\Big] &= \sum_i \sum_j \Big[\sigma^2 - \frac{2\sigma^2}{n} + \frac{\sigma^2}{n}\Big] \\ &= \sum_i \sum_j \frac{n-1}{n}\sigma^2 \\ &= (n-1)\sigma^2 \end{aligned}$$

Hence $\frac{1}{n-1}SSE_0 = \frac{1}{n-1}\sum_i \sum_j \Big(Y_{ij} - \bar{Y}_{\cdot\cdot}\Big)^2$ would be an unbiased estimator for σ^2

Question 6.

Solution 6. *Notice the simple fact that for any column vector $a = (a_1, a_2, \ldots, a_n)^T$, we have:*

$$a^T J a = \Big(\sum_{i=1}^n a_i\Big)^2$$
$$a^T I a = \sum_{i=1}^n a_i^2$$

where J is an $n \times n$ matrix whose every entry is 1, and I is an $n \times n$ identity matrix.

Now let's first look at SSTOT:

$$\begin{aligned} SSTOT &= \sum\sum \Big(Y_{ij} - \bar{Y}_{\cdot\cdot}\Big)^2 \\ &= \sum\sum \Big(Y_{ij}^2 - 2Y_{ij}\bar{Y}_{\cdot\cdot} + \bar{Y}_{\cdot\cdot}^2\Big) \\ &= \sum\sum Y_{ij}^2 - 2\bar{Y}_{\cdot\cdot}\sum\sum Y_{ij} + \sum\sum \bar{Y}_{\cdot\cdot}^2 \\ &= \sum\sum Y_{ij}^2 - \frac{2}{N}\Big(\sum\sum Y_{ij}\Big)^2 + \frac{1}{N}\Big(\sum\sum Y_{ij}\Big)^2 \\ &= \sum\sum Y_{ij}^2 - \frac{1}{N}\Big(\sum\sum Y_{ij}\Big)^2 \\ &= Y^T I Y - \frac{1}{N} Y^T J Y \\ &= Y^T\Big(I - \frac{1}{N}J\Big)Y \end{aligned}$$

Here I is an $N \times N$ identity matrix and J is an $N \times N$ matrix all those entries are 1. Also $Y = \Big(Y_{11}, \ldots, Y_{1n_1}, \ldots, Y_{g1}, \ldots, Y_{gn_g}\Big)^T$ is a column vector whose length is $\sum_{i=1}^g \sum_{j=1}^{n_i} 1 = N$.

Similarly, for error sum of square:

$$\begin{aligned}
SSE &= \sum\sum\Big(Y_{ij} - \bar{Y}_{i\cdot}\Big)^2 \\
&= \sum\sum\Big(Y_{ij}^2 - 2Y_{ij}\bar{Y}_{i\cdot} + \bar{Y}_{i\cdot}^2\Big) \\
&= \sum\sum Y_{ij}^2 - 2\sum\sum Y_{ij}\bar{Y}_{i\cdot} + \sum\sum \bar{Y}_{i}^2 \\
&= \sum\sum Y_{ij}^2 - 2\sum \bar{Y}_{i\cdot}\Big(\sum Y_{ij}\Big) + \sum n_i\bar{Y}_{i\cdot}^2 \\
&= \sum\sum Y_{ij}^2 - 2\sum \frac{1}{n_i}\Big(\sum Y_{ij}\Big)^2 + \sum \frac{1}{n_i}\Big(\sum Y_{ij}\Big)^2 \\
&= \sum\sum Y_{ij}^2 - \sum \frac{1}{n_i}\Big(\sum Y_{ij}\Big)^2 \\
&= Y^TIY - \sum \frac{1}{n_i}Y_i^TJ_{n_i}Y_i \text{ (I will explain the notation below)} \\
&= Y^TIY - Y^T\begin{bmatrix} \frac{1}{n_1}Y_{n_1}^TJ_{n_1}Y_{n_1} & & \\ & \ddots & \\ & & \frac{1}{n_g}Y_{n_g}^TJ_{n_g}Y_{n_g} \end{bmatrix}Y \\
&= Y^TIY - Y^THY \\
&= Y^T\Big(I - H\Big)Y
\end{aligned}$$

Here

$$H = \begin{bmatrix} \frac{1}{n_1}Y_{n_1}^TJ_{n_1}Y_{n_1} & \cdots & \cdots \\ \vdots & \ddots & \vdots \\ \cdots & \cdots & \frac{1}{n_g}Y_{n_g}^TJ_{n_g}Y_{n_g} \end{bmatrix}$$

Also, $Y_{n_i} = \Big(Y_{i1}, Y_{i2}, \ldots, Y_{in_i}\Big)^T$ *is a column vector indicating the* $i-th$ *treatment with* n_i *observations.* J_{n_i} *is an* $n_i \times n_i$ *matrix all whose entries are* 1.

Finally for SST, it is well know that

$$\begin{aligned}
SST &= SSTOT - SSE \\
&= Y^T\Big(I - \frac{1}{N}J\Big)Y - Y^T\Big(I - H\Big)Y \\
&= Y^T\Big(I - \frac{1}{N}J - I + H\Big)Y \\
&= Y^T\Big(H - \frac{1}{N}J\Big)Y
\end{aligned}$$

Thus finished the proof.

Question 7.

Solution 7. *For part* (a)*:*

Since $Y_{ij}'s$ are independent and normal, we know that each $\bar{Y}_{i\cdot}$ is normal for any $i = 1, 2, \ldots, g$. Meanwhile $\bar{Y}_{i\cdot}, i = 1, 2, \ldots, g$ are independent.

We have:

$$\begin{aligned} E\Big[\bar{Y}_{i\cdot}\Big] &= \frac{1}{n_i}\sum_{i=1}^{n_i} E\Big[Y_{ij}\Big] \\ &= \frac{1}{n_i}\sum_{i=1}^{n_i}\mu_i \\ &= \frac{1}{n_i}\cdot n_i\mu_i \\ &= \mu_i \\ Var\Big[\bar{Y}_{i\cdot}\Big] &= \frac{n_i\sigma^2}{n_i^2} = \frac{\sigma^2}{n_i} \end{aligned}$$

So $\bar{Y}_{i\cdot} \sim N\Big(\mu_i, \frac{\sigma^2}{n_i}\Big)$.

Due to the independence among $\bar{Y}_{i\cdot}$, we know that $\sum c_i\bar{Y}_{i\cdot}$ is also normal. We have:

$$\begin{aligned} E\Big[\sum c_i\bar{Y}_{i\cdot}\Big] &= \sum c_i E\Big[\bar{Y}_{i\cdot}\Big] = \sum c_i\mu_i \\ Var\Big[\sum c_i\bar{Y}_{i\cdot}\Big] &= \sum c_i^2\, Var\Big[\bar{Y}_{i\cdot}\Big] = \sum c_i^2\cdot\frac{\sigma^2}{n_i} = \sigma^2\sum\frac{c_i^2}{n_i} \end{aligned}$$

So

$$\sum c_i\bar{Y}_{i\cdot} \sim N\Big(\sum c_i\mu_i, \sigma^2\sum\frac{c_i^2}{n_i}\Big)$$

and hence

$$X = \frac{\sum c_i\bar{Y}_{i\cdot} - \sum c_i\mu_i}{\sqrt{\sigma^2\sum c_i^2/n_i}} \sim N\Big(0,1\Big)$$

For part (b):

Within each treatment group $i, 1 \le i \le g$, we notice that:

$$\sum_{j=1}^{n_i}\frac{\Big(Y_{ij}-\bar{Y}_{i\cdot}\Big)^2}{\sigma^2} = \frac{(n_i-1)\cdot\frac{1}{n_i-1}\sum_{j=1}^{n_i}\Big(Y_{ij}-\bar{Y}_{i\cdot}\Big)^2}{\sigma^2} = \frac{(n_i-1)S_i^2}{\sigma^2} \sim \chi^2_{n_i-1}$$

where S_i^2 is the sample variance of treatment group i.

Notice that across the groups $\sum_{j=1}^{n_i}\frac{\Big(Y_{ij}-\bar{Y}_{i\cdot}\Big)^2}{\sigma^2}$ are independent due to the model assumption, so

$$\frac{SSE}{\sigma^2} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}\frac{\Big(Y_{ij}-\bar{Y}_{i\cdot}\Big)^2}{\sigma^2} = \sum_{i=1}^{g}\frac{(n_i-1)S_i^2}{\sigma^2} \sim \chi^2_{\sum_{i=1}^{g}(n_i-1)} = \chi^2_{n-g}$$

Question 8.

Solution 8. *For part (a):*

Let's first compute the varaince of $\bar{X}$:

Under null hypothesis and assumption of the model:

$$\begin{aligned} Var(\bar{X}) &= E\Big[\Big(\bar{X}\Big)^2\Big] = \frac{1}{n^2}E\Big[\Big(\sum X_i\Big)^2\Big] \\ &= \frac{1}{n^2}\Big(E\Big[\sum X_i^2\Big] + E\Big(\sum_{k=1}^{n-1}\sum_{|i=j|=k} E\Big(X_iX_j\Big)\Big)\Big) \\ &= \frac{1}{n^2}\Big(n\sigma^2 + \sum_{k=1}^{n-1}\sum_{|i-j|=k}\sigma^2\cdot\rho^k\Big) \end{aligned}$$

The last step is due to: when $|i-j|=k$,

$$\begin{aligned} \rho^k &= \frac{Cov(X_i,X_j)}{\sqrt{Var(X_i)}\sqrt{Var(X_j)}} \\ &= \frac{E(X_iX_j) - \cancel{E(X_i)E(X_j)}}{\sigma^2} \end{aligned}$$

So

$$E(X_iX_j) = \sigma^2\rho^k$$

Notice that for each case $|i-j|=k$, there are 2 terms satisfy this for same pair of (i,j). There are also $n-k$ pairs of different (i,j) such that $|i-j|=k$. So continue from above, we have:

$$\begin{aligned} Var(\bar{X}) &= \frac{1}{n^2}\Big(n\sigma^2 + 2\sum_{k=1}^{n-1}\sigma^2\rho^k(n-k)\Big) \\ &= \frac{1}{n^2}\Big[n\sigma^2 + 2\sigma^2\sum_{k=1}^{n-1}(n-k)\rho^k\Big] \\ &= \frac{1}{n^2}\Big[n\sigma^2 + 2\sigma^2 n\sum_{k=1}^{n-1}\rho^k - 2\sigma^2\sum_{k=1}^{n-1}k\rho^k\Big] \\ &= \frac{1}{n^2}\Big[n\sigma^2 + 2n\sigma^2\cdot\frac{\rho-\rho^n}{1-\rho} - 2\sigma^2\rho\sum_{k=1}^{n-1}k\cdot\rho^{k-1}\Big] \\ &= \frac{1}{n^2}\Big[n\sigma^2 + 2n\sigma^2\cdot\frac{\rho-\rho^n}{1-\rho} - 2\sigma^2\rho\cdot\Big(\sum_{k=1}^{n-1}\rho^k\Big)'\Big] \\ &= \frac{1}{n^2}\Big[n\sigma^2 + 2n\sigma^2\cdot\frac{\rho-\rho^n}{1-\rho} - 2\sigma^2\rho\cdot\frac{(1-n\rho^{n-1})(1-\rho)+(\rho-\rho^n)}{(1-\rho)^2}\Big] \end{aligned}$$

Continue from above, we got:

$$\begin{aligned} Var(\bar{X}) &= \frac{\sigma^2}{n}\Big[\frac{(1-\rho)^2 + 2(\rho-\rho^n)(1-\rho) - \frac{2}{n}\rho\Big(1-\cancel{\rho} - n\rho^{n-1} + n\rho^n + \cancel{\rho} - \rho^n\Big)}{(1-\rho)^2}\Big] \\ &= \frac{\sigma^2}{n}\cdot\frac{1-\cancel{2\rho}+\rho^2+\cancel{2\rho}-2\rho^2-\cancel{2\rho^n}+\cancel{2\rho^{n+1}}-\frac{2}{n}\rho+\cancel{2\rho^n}-\cancel{2\rho^{n+1}}+\frac{2}{n}\rho^{n+1}}{(1-\rho)^2} \\ &= \frac{\sigma^2}{n}\cdot\frac{1-\rho^2-\frac{2}{n}\rho+\frac{2}{n}\rho^{n+1}}{(1-\rho)^2} \end{aligned}$$

When n is large enough, we could give up the two terms $\frac{2}{n}\rho$ and $\frac{2}{n}\rho^{n+1}$, so

$$Var(\bar{X}) \simeq \frac{\sigma^2}{n}\cdot\frac{1-\rho^2}{(1-\rho)^2} = \frac{\sigma^2(1+\rho)}{n(1-\rho)}$$

take square root, we have:

$$\sigma(\bar{X}) \simeq \sqrt{\frac{1+\rho}{1-\rho}}\cdot\frac{\sigma}{\sqrt{n}}$$

For part (b):

If we use one sample $t-test$, then our statistic would be:

$$t = \frac{\bar{X}}{\frac{s}{\sqrt{n}}\cdot\sqrt{\frac{1+\rho}{1-\rho}}}$$

When $\rho > 0$, we have $\frac{1+\rho}{1-\rho} > 1$, thus we got smaller absolute value for our t-statistic, which means our p value is larger, and we reject less. Thus the chance of making Type I error (falsely reject) will be smaller.

For part (c):

Similar to (b) except this is the oppositve case. When $\rho < 0$, $\frac{1+\rho}{1-\rho} < 1$, so our t statistic has larger absolute value, hence we got smaller p value and reject more often, hence increase the chance of making type I error.

Question 9.

Solution 9. *The code is attached as HW1Q9.R. I am going to explain the process of the code here:*

1. *with each given ρ, we sample 1000 times from $N(0, 25)$, each sample is of size $10,000$. We have to be careful here because within each sample, the sample points are not independent. We use an $AR(1)$ time series model to simulate our data.*

2. *run one sample t-test(with modifided statistic involving ρ) on all 1000 sample sets for each ρ value. Since we have prior knowledge that it is sampled from $N(0, 25)$, so the null hypothesis $\mu = 0$ should be true. Thus any rejection would be recorded as a type I error. We record the type I error rate under each given value ρ.*

3. compare our results with our theoretical expectation from Question 8.

Also, we need to write our own one sample t-test function, because the default t-test given by R is only working for independent samples.

After we run the code, we got the following results:

ρ	$\sqrt{\frac{1+\rho}{1-\rho}}$	*Type I error*
-0.50	*0.58*	*0.045*
-0.25	*0.77*	*0.055*
-0.10	*0.90*	*0.053*
0.00	*1.00*	*0.052*
0.10	*1.11*	*0.045*
0.25	*1.29*	*0.039*
0.50	*1.73*	*0.046*

It is clear to see that from $\rho[2]$ to $\rho[6]$ as the value of ρ increase from negative to positive, the type I error rate decrease, which match with conclusion from Question 8. However $\rho[1]$ and $\rho[7]$ behaves abnormaly. I think my code is somewhere problamatic still, because even I can run though it, I will always get warning. Need to further check on this.