Problem 1.

a) $E[\bar{X}] = E[\frac{1}{n}\sum_{i=1}^{n} X_i]$

$= \frac{1}{n} E[\sum_{i=1}^{n} X_i]$

$= \frac{1}{n}\sum_{i=1}^{n} E[X_i]$

$\because X_i \sim N(\mu, \sigma^2)$

$\therefore E[X_i] = \mu$

$\therefore E[\bar{x}] = \frac{1}{n} \cdot n\cdot\mu = \mu \Rightarrow \bar{x}$ is an unbiased estimator of μ

b) $E[S^2] = E[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2]$

$= \frac{1}{n-1} E[\sum_{i=1}^{n}(X_i - \bar{X})^2]$

$= \frac{1}{n-1} E[\sum_{i=1}^{n} X_i^2 + \sum_{i=1}^{n}\bar{X}^2 - \sum_{i=1}^{n} 2X_i\bar{X}]$

$= \frac{1}{n-1}[\sum_{i=1}^{n} E[X_i^2] + E[n\bar{X}^2 - 2n\bar{X}^2]]$

$= \frac{1}{n-1}[\sum_{i=1}^{n} E[X_i^2] - E[n\bar{X}^2]]$

$= \frac{n}{n-1}\cdot(\frac{1}{n}\sum E(X_i^2) - E[\bar{X}^2])$ ①

$E[X_i^2] = Var[X_i] + E^2[X_i] = \sigma^2 + \mu^2$

$E[\bar{X}^2] = Var[\bar{X}] + E^2[\bar{X}] = \frac{\sigma^2}{n} + \mu^2$

$\therefore$ ① $= \frac{n}{n-1}(\sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2)$

$= \frac{n}{n-1}\cdot\sigma^2\cdot\frac{n-1}{n}$

$= \sigma^2$

$\therefore E[S^2] = \sigma^2 \Rightarrow S^2$ is an unbiased estimator of σ^2

c) $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}[(y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y})]^2$

$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i - \bar{y})^2 + \underbrace{2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y})}_{①}$

To prove the partitioning of the total variability

$\Rightarrow$ to prove that ① $= 0$:

$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y}) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}\bar{y}_i - y_{ij}\bar{y} - \bar{y}_i^2 + \bar{y}_i\bar{y}) = \sum_{i=1}^{k}\bar{y}_i\cdot n_i\bar{y}_i - \bar{y}\cdot n\bar{y} - \sum_{i=1}^{k} n_i\bar{y}_i^2 + \sum_{i=1}^{k}\bar{y}n_i\bar{y}_i$

$= \sum_{i=1}^{k} n_i\bar{y}_i^2 - \sum_{i=1}^{k} n_i\bar{y}_i^2 - n\bar{y}^2 + \bar{y}\sum_{i=1}^{k} n_i\bar{y}_i$

$= -n\bar{y}^2 + \bar{y}\cdot n\cdot\bar{y}$

$= 0$ □ $\Rightarrow \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i - \bar{y})^2$

p8130_hw3_xj2249

xj2249

2019/10/24

Problem2

a)

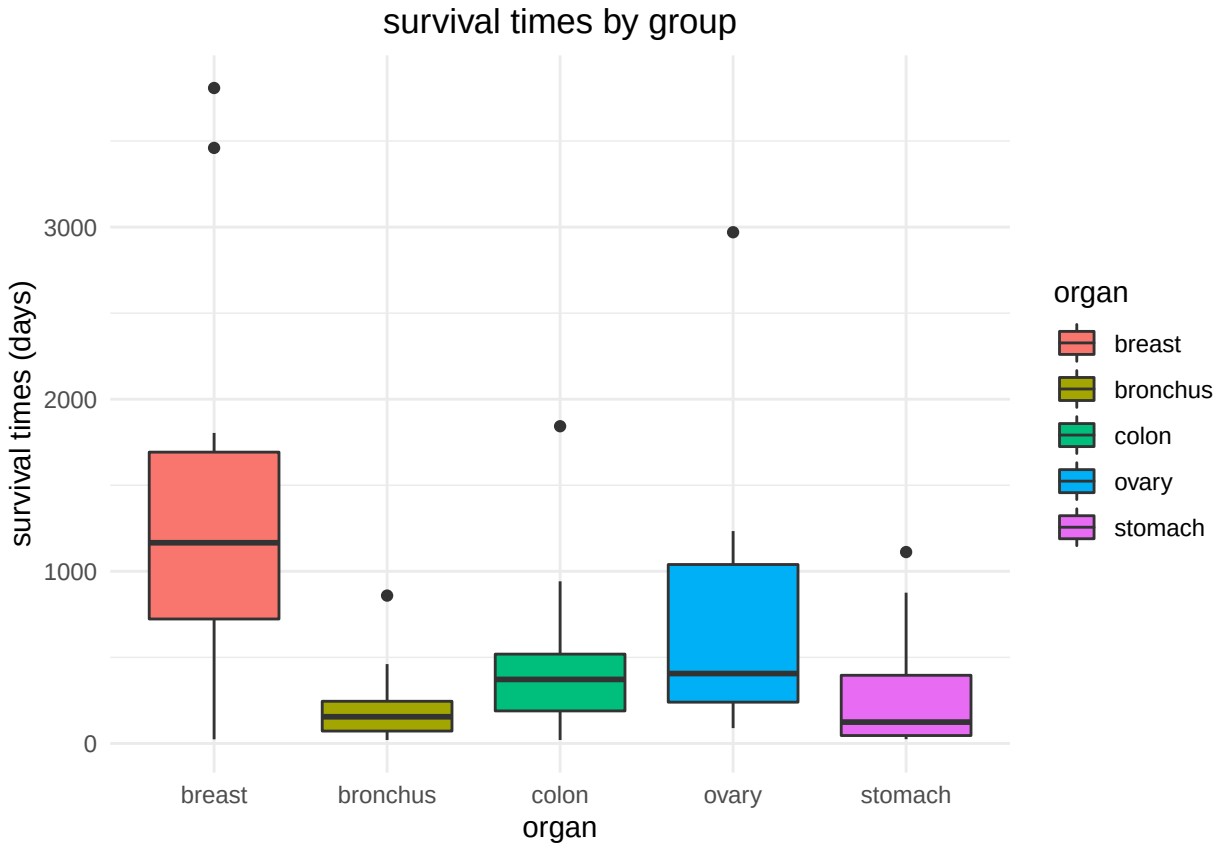
```
# descriptive statistics
control_table1 <- tableby.control(
        total = FALSE,
        test = FALSE,
        numeric.stats = c("meansd","medianq1q3","range"),
        stats.labels = list(meansd = "Mean (SD)",
                            medianq1q3 = "Median (Q1, Q3)",
                            range = "Min - Max"),
        digits = 2
        )
tab1 <- tableby(organ~survival,sur_df,control = control_table1)

summary(tab1,text = TRUE) %>% kable(caption = "Descriptive statistics for each group") %>% kable_stylin
```

```
sur_df %>%
        ggplot(aes( x =  organ, y = survival , fill = organ)) +
        geom_boxplot() +
        labs( y = "survival times (days)",
              title = "survival times by group")
```

Table 1: Descriptive statistics for each group

	breast (N=11)	bronchus (N=17)	colon (N=17)	ovary (N=6)	stomach (N=13)
survival					
- Mean (SD)	1395.91 (1238.97)	211.59 (209.86)	457.41 (427.17)	884.33 (1098.58)	286.00 (346.31)
- Median (Q1, Q3)	1166.00 (723.00, 1692.50)	155.00 (72.00, 245.00)	372.00 (189.00, 519.00)	406.00 (239.75, 1039.50)	124.00 (46.00, 396.00)
- Min - Max	24.00 - 3808.00	20.00 - 859.00	20.00 - 1843.00	89.00 - 2970.00	25.00 - 1112.00



As we can see, average survival time varies among different cancers. Breast cancer group has the largest mean and standard deviation value, followed by ovary, colon, stomach and bronchus cancer groups.

b)

Table 2: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
organ	4	11535761	2883940	6.433	0.0002295
Residuals	59	26448144	448274	NA	NA

1) Hypothesis:
$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 \quad vs \quad H_1 : \text{not all means are equal}$$

2) Significance level: $\alpha = 0.01$

3) Assumptions:(i)Independence (ii)equal variances (iii)normality

4) Decision rule:
Reject H_0: if $F_{stats} > F_{4,59,1-\alpha} = F_{4,59,0.99} = 3.655$
Fail to reject H_0: if $F_{stats} < F_{4,59,0.99}$

5) Interpretation: Since $F_{stats} = 6.433 > F_{4,59,1-\alpha} = F_{4,59,0.99} = 3.655$, we reject H_0 and conclude that there is a significant difference in average survival time among different cancer groups.

c) pairwise comparisons

1) Bonferroni

```
pairwise.t.test(sur_df$survival, sur_df$organ, p.adj = 'bonferroni', conf.level = 0.99)
```

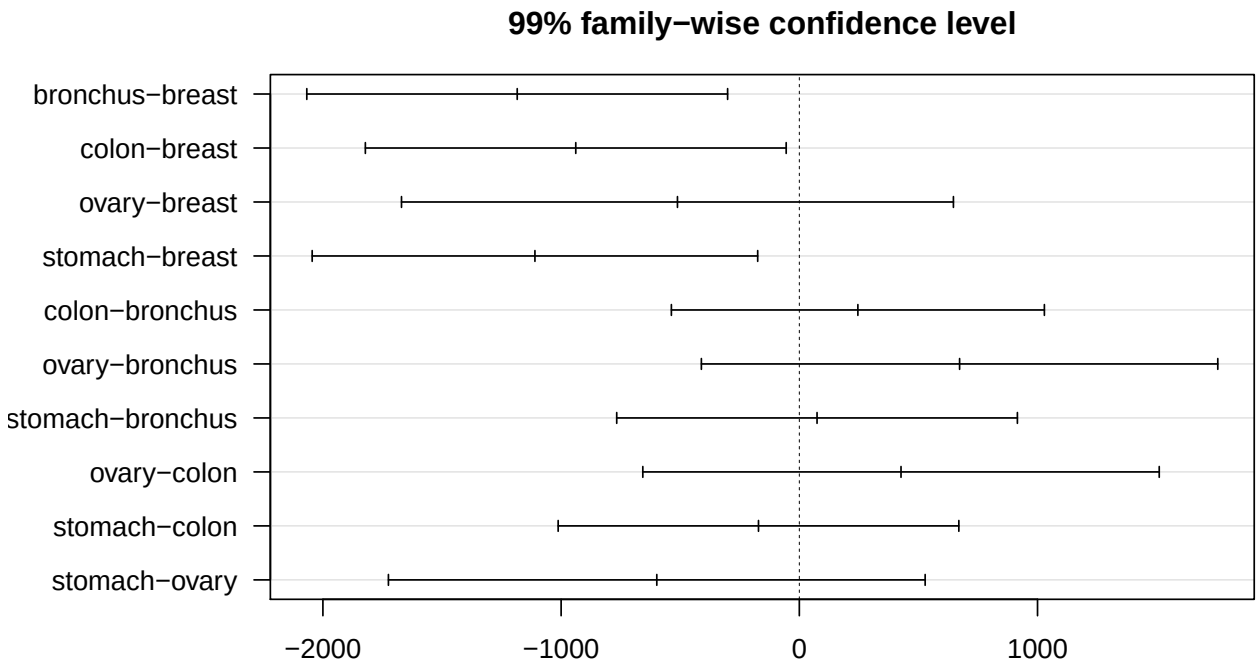
```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  sur_df$survival and sur_df$organ
##
##          breast  bronchus colon   ovary
## bronchus 0.00025 -        -       -
## colon    0.00608 1.00000  -       -
## ovary    1.00000 0.38575  1.00000 -
## stomach  0.00153 1.00000  1.00000 0.75283
##
## P value adjustment method: bonferroni
```

2) Tukey

```
TukeyHSD(sur_aov,conf.level = 0.99)
```

```
##   Tukey multiple comparisons of means
##     99% family-wise confidence level
##
## Fit: aov(formula = survival ~ organ, data = sur_df)
##
## $organ
##                          diff        lwr        upr     p adj
## bronchus-breast  -1184.32086 -2067.6073 -301.03446 0.0002385
## colon-breast      -938.49733 -1821.7837  -55.21093 0.0053072
## ovary-breast      -511.57576 -1670.0752  646.92367 0.5630900
## stomach-breast   -1109.90909 -2045.0583 -174.75983 0.0013962
## colon-bronchus     245.82353  -537.1262 1028.77324 0.8208402
## ovary-bronchus     672.74510  -411.1997 1756.68989 0.2271084
## stomach-bronchus    74.41176  -766.6111  915.43467 0.9981461
## ovary-colon        426.92157  -657.0232 1510.86636 0.6659115
## stomach-colon     -171.41176 -1012.4347  669.61114 0.9568289
## stomach-ovary     -598.33333 -1724.9413  528.27467 0.3772923
```

```
par(mar=c(2,8,2,2))
TukeyHSD(sur_aov,conf.level = 0.99) %>% plot(las = 1)
```



3) Dunnett Test

```
glht(sur_aov, linfct = mcp(organ = "Dunnett"),conf.level =0.99) %>% summary()
```

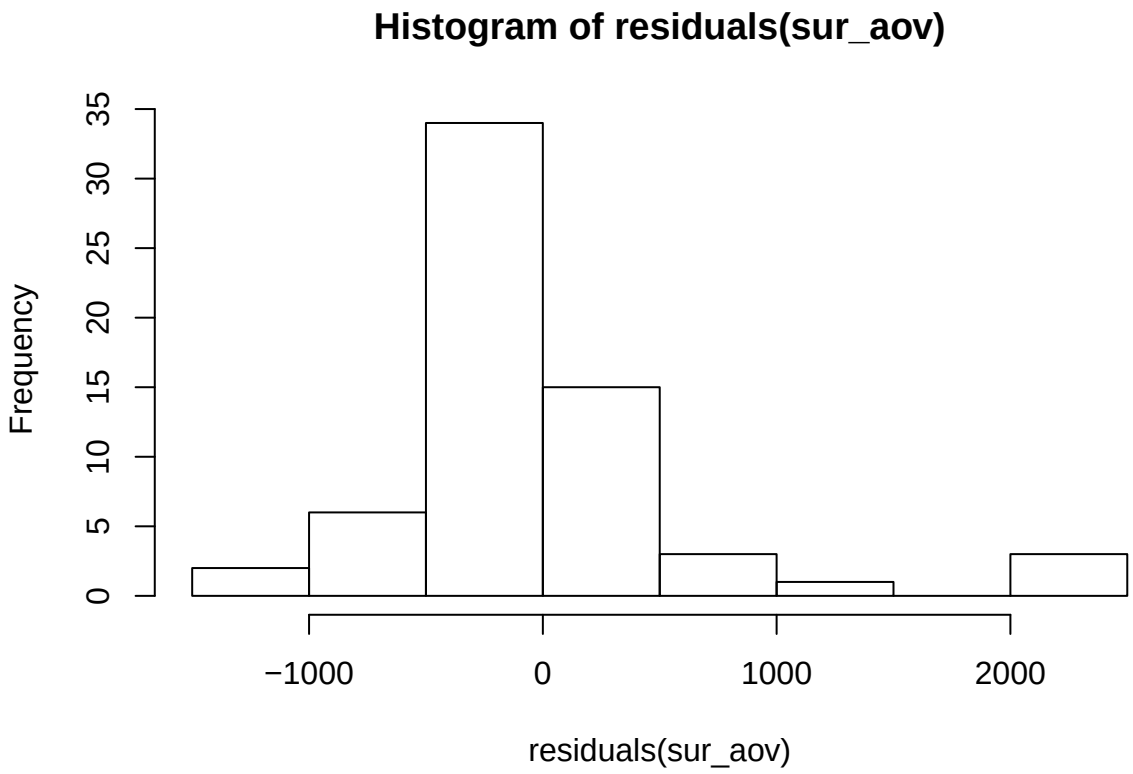
```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Dunnett Contrasts
##
##
## Fit: aov(formula = survival ~ organ, data = sur_df)
##
## Linear Hypotheses:
##                          Estimate Std. Error t value Pr(>|t|)
## bronchus - breast == 0  -1184.3      259.1  -4.571  < 0.001 ***
## colon - breast == 0      -938.5      259.1  -3.622  0.00228 **
## ovary - breast == 0      -511.6      339.8  -1.506  0.36692
## stomach - breast == 0   -1109.9      274.3  -4.046  < 0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

The main conclusions of these three methods are consistent with each other: at a significance level of 0.01, we can conclude that there's a significant difference in average survival time between bronchus and breast cancer; colon and breast cancer; stomach and breast cancer. According to Bonferroni and Tukey, we can also conclude that there's no there's no enough evidence to support a significant difference between other pairwise groups.
As we can see, the p-value using Tukey's method is smaller than that of Bonferroni's methods, indicating than Tukey is less conservative than Bonferroni. And Dunnett is used to compare with a specific group rather than any pairwise comparisons, like Bonferroni and Tukey.

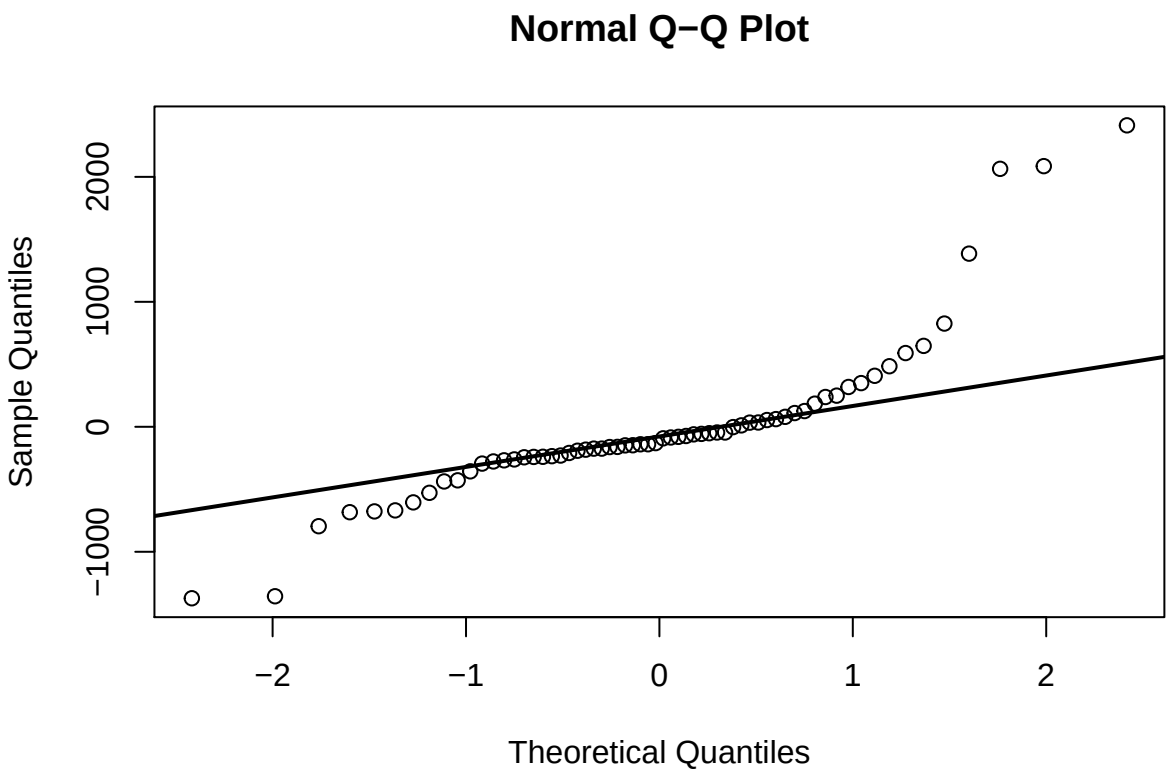
d)

i) check the normality assumption

```
# first, try a hist/density plot.
hist(residuals(sur_aov))
```




```
# check the normality (of residuals) assumption
qqnorm(residuals(sur_aov))
qqline(residuals(sur_aov),lwd = 2)
```



As the QQ-plot shows, the normality assumption is questionable. Therefore, we can use **non-parametric test (Kruskal Wallis test)** or **transformation** to fix the problem.

ii) KW test

```
kruskal.test(survival ~ organ, data = sur_df) %>% pander::pander()
```

Table 3: Kruskal-Wallis rank sum test: `survival` by `organ`

Test statistic	df	P value
14.95	4	0.004798 * *

Since p-value < 0.01, at a significance level 0.01, we reject H_0 and conclude that there is a significant difference in average survival time among different cancer groups. The p-value of kw-test is 0.004798, much larger than that of the anova test(0.0002295), which shows that kw-test is harder to reject H_0 and it's more conservative and less powerful.

Problem3

a)

$$\begin{aligned} Average_{zinc} &= \frac{n_{zinc} \times Avg_{zinc} + n_{zinc+iron} \times Avg_{zinc+iron}}{n_{iron} + n_{zinc+iron}} \\ &= \frac{54 \times 0.7 + 55 \times 0.8}{54 + 55} \\ &= 0.75 \end{aligned}$$

$$
\begin{aligned}
Average_{non-zinc} &= \frac{n_{placebo} \times Avg_{placebo} + n_{iron} \times Avg_{iron}}{n_{placebo} + n_{iron}} \\
&= \frac{56 \times 1.1 + 54 \times 1.4}{56 + 54} \\
&= 1.25
\end{aligned}
$$

$$
diff = Average_{zinc} - Average_{non-zinc} = -0.5
$$

b)

$$
\begin{aligned}
sd_{placebo} &= s.e._{placebo} \times \sqrt{n_{placebo}} = \sqrt{56} \times 0.2 = 1.5 \\
sd_{iron} &= s.e._{iron} \times \sqrt{n_{iron}} = \sqrt{54} \times 0.2 = 1.47 \\
sd_{zinc} &= s.e._{zinc} \times \sqrt{n_{zinc}} = \sqrt{54} \times 0.1 = 0.73 \\
sd_{zinc+iron} &= s.e._{zinc+iron} \times \sqrt{n_{zinc+iron}} = \sqrt{55} \times 0.1 = 0.74
\end{aligned}
$$

Between 4 initial groups, placebo and iron group share similar and larger standard deviation, zinc and zinc+iron group share similar and smaller standard deviation.

To decide if we can assume equal variances, we can use F-test. For "zinc" and "zinc-iron", $F_{stats} = \frac{s^2_{placebo}}{s^2_{zinc+iron}} = \frac{0.73^2}{0.74^2} = 0.97 \leq F_{53,54,1-0.05/2} = 1.72$ and $\geq F_{53,54,0.05/2} = 0.58$. For "placebo" and "iron" groups, $F_{stats} = \frac{s^2_{zinc}}{s^2_{iron}} = \frac{1.50^2}{1.47^2} = 1.04 \leq F_{55,53,1-0.05/2} = 1.72$ and $\geq F_{55,53,0.05/2} = 0.58$

Therefore, we can assume that the standard deviation is equal b/w "zinc" and "zinc-iron", and also b/w "placebo" and "iron".

$$
\begin{aligned}
s^2_{zinc} &= \frac{s^2_{zinc} \times (n_{zinc} - 1) + s^2_{zinc+iron} \times (n_{zinc+iron} - 1)}{n_{iron} + n_{zinc+iron} - 2} \\
&= \frac{0.73^2 \times 53 + 0.74^2 \times 54}{54 + 55 - 2} \\
&= 0.54 \\
s^2_{non-iron} &= \frac{s^2_{placebo} \times (n_{placebo} - 1) + s^2_{iron} \times (n_{iron} - 1)}{n_{placebo} + n_{iron} - 2} \\
&= \frac{1.50^2 \times 55 + 1.47^2 \times 53}{56 + 54 - 2} \\
&= 2.21
\end{aligned}
$$

c)

d) Equal allocation

$$
\begin{aligned}
n &= \frac{\left(\sigma^2_{zinc} + \sigma^2_{non-zinc}\right)\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\Delta^2} \\
&= \frac{(0.54 + 2.21)(1.96 + 0.84)^2}{0.5^2} \\
&= 86.24 \\
&\approx 87
\end{aligned}
$$

Therefore, the sample size is 87 for both zinc and non-zinc group.

ii) 2:1 allocation

$$
k = n_{non-zinc}/n_{zinc} = 2
$$

$$
\begin{aligned}
n_{zinc} &= \frac{\left(\sigma^2_{zinc} + \sigma^2_{non-zinc}/k\right)\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\Delta^2} \\
&= \frac{(0.54 + 2.21/2)(1.96 + 0.84)^2}{0.5^2} \\
&= 51.59 \\
&\approx 52
\end{aligned}
$$

Therefore, the sample size for zinc group is 52.

$$
\begin{aligned}
n_{non-zinc} &= \frac{\left(k\sigma^2_{zinc} + \sigma^2_{non-zinc}\right)\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\Delta^2} \\
&= \frac{(2 \times 0.54 + 2.21)(1.96 + 0.84)^2}{0.5^2} \\
&= 103.17 \\
&\approx 104
\end{aligned}
$$

Therefore, the sample size for non-zinc group is 104.