BCB744 (BioStatistics): Summative Task 2, 12 April 2023

AJ Smit

2023-05-02

On this page

0.1 Honesty Pledge

This assignment requires that you work as an individual and not share your code, results, or discussion with your peers. Penalties and disciplinary action will apply if you are found cheating.

Acknowledgement of the Pledge

Copy the statement, below, into your document and replace the underscores with your name acknowledging adherence to the UWC's Honesty Pledge.
I, _____, hereby state that I have not communicated with or gained information in any way from my peers and that all work is my own.

0.2 Instructions

Please note the following instructions. Failing to comply with them in full will result in a loss of marks.

- **QUARTO –> HTML** Submit your assessment answers as an .html file compiled from your Quarto document. Produce *fully annotated reports*, including the meta-information at the top (name, date, purpose, etc.). Provide ample commentary explaining the purpose of the various tests/sections as necessary.
- **TESTING OF ASSUMPTIONS** For all questions, make sure that when *formal inferential statistics are required, each is preceded by the appropriate tests for the assumptions*, i.e., state the assumptions, state the statistical procedure for testing the assumptions and mention their corresponding H_0. If a graphical approach is used to test assumptions, explain the principle behind the approach. Explain the findings emerging from the test of assumptions, and justify your selection of the appropriate inferential test (e.g. *t*-test, ANOVA, etc.) that you will use.
- **STATE HYPOTHESES** When inferential statistics are required, please provide the full H_0 and H_A, and conclude the analysis with a statement of which is accepted or rejected.
- **GRAPHICAL SUPPORT** All descriptive and inferential statistics must be supported by the appropriate figures of the results.
- **STATEMENT OF RESULTS** Make sure that the textual statement of the final result is written exactly as required for it to be published in a journal article. Please consult a journal if you don't know how.
- **FORMATTING** Pay attention to formatting. Some marks will be allocated to the appearance of the script, including considerations of aspects of the tidiness of the file, the use of the appropriate headings, and adherence to code conventions (e.g. spacing etc.).
- **MARK ALLOCATION** Please see the Introduction Page for an explanation of the assessment approach that will be applied to these questions.

Submit the .html file wherein you provide answers to Questions 1–7 by no later than 19:00 today. Label the script as follows:

BCB744_<Name>_<Surname>_Summative_Task_2.html, e.g.

BCB744_AJ_Smit_Summative_Task_2.html.

Upload your .html files onto Google Forms.

0.3 Question 1

0.3.1 Chromosomal effects of mercury-contaminated fish consumption

These data reside in package **coin**, dataset `mercuryfish`. The dataframe contains the mercury level in blood, the proportion of cells with abnormalities, and the proportion of cells with chromosome aberrations in consumers of mercury-contaminated fish and a control group. Please see the dataset's help file for more information.

Analyse the dataset and answer the following questions:

a. Does the presence of methyl-mercury in a diet containing fish result in a higher proportion of cellular abnormalities?

b. Does the concentration of mercury in the blood influence the proportion of cells with abnormalities, and does this differ between the `control` and `exposed` groups?

c. Is there a relationship between the variables `abnormal` and `ccells`? This will have to be for the `control` and `exposed` groups, noting that an interaction effect *might* be present.

0.4 Question 2

0.4.1 Malignant glioma pilot study

Package **coin**, dataset `glioma`: A non-randomized pilot study on malignant glioma patients with pretargeted adjuvant radioimmunotherapy using yttrium-90-biotin.

a. Do `sex` and `group` interact to affect survival time (`time`)?

b. Do `age` and `histology` interact to affect survival time (`time`)?

c. Show a full graphical exploration of the data. Are there any other remaining patterns visible in the data that should be explored statistically? Study your results, select the most promising and insightful question that remains, and do the analysis.

0.5 Question 3

0.5.1 Risk factors associated with low infant birth weight

Package **MASS**, dataset `birthwt`: A dataset about the risk factors associated with low infant birth mass collected at Baystate Medical Center, Springfield, Mass. during 1986.

State three hypotheses and test them. Make sure one of the tests makes use of the 95% confidence interval approach rather than a formal inferential methodology.

0.6 Question 4

0.6.1 The LungCapData.csv data

a. Using the Lung Capacity data provided, please calculate the 95% CIs for the `LungCap` variable as a function of:

- `Gender`
- `Smoke`
- `Caesarean`

```
lungs <- read.csv("../data/LungCapData.csv", sep = "\t")

library(rcompanion)

(gender_ci <- groupwiseMean(LungCap ~ Gender, data = lungs, conf = 0.95, digits = 3))
```

```
  Gender   n Mean Conf.level Trad.lower Trad.upper
1 female 358 7.41       0.95       7.14       7.67
2   male 367 8.31       0.95       8.03       8.58
```

```
(smoke_ci <- groupwiseMean(LungCap ~ Smoke, data = lungs, conf = 0.95, digits = 3))
```

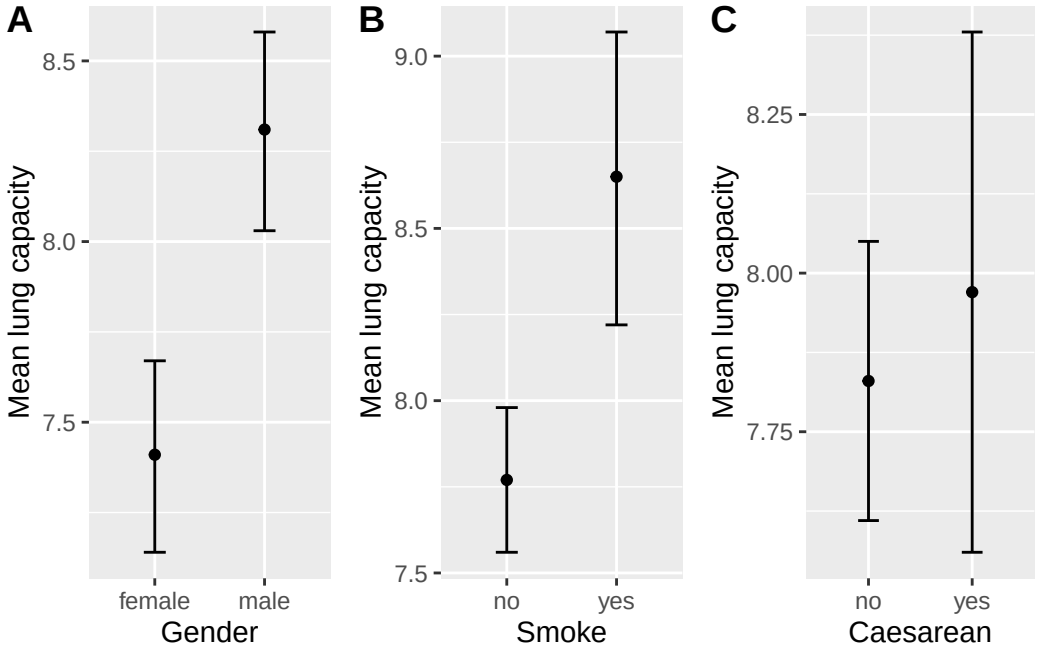
```
  Smoke   n Mean Conf.level Trad.lower Trad.upper
1    no 648 7.77       0.95       7.56       7.98
2   yes  77 8.65       0.95       8.22       9.07
```

```
(caesarean_ci <- groupwiseMean(LungCap ~ Caesarean, data = lungs, conf = 0.95, digits = 3)
```

```
  Caesarean   n Mean Conf.level Trad.lower Trad.upper
1        no 561 7.83       0.95       7.61       8.05
2       yes 164 7.97       0.95       7.56       8.38
```

b. Create a graph of the mean ± 95% CIs and determine if there are statistical differences in `LungCap` between the levels of `Gender`, `Smoke`, and `Caesarean`. Do the same using inferential statistics. Are your findings the same using these two approaches?

```
plt1 <- ggplot(gender_ci, aes(x = Gender, y = Mean)) +
  geom_point() +
  geom_errorbar(aes(ymin = Trad.lower, ymax = Trad.upper), width = 0.2) +
  ylab("Mean lung capacity")

plt2 <- ggplot(smoke_ci, aes(x = Smoke, y = Mean)) +
  geom_point() +
  geom_errorbar(aes(ymin = Trad.lower, ymax = Trad.upper), width = 0.2) +
  ylab("Mean lung capacity")

plt3 <- ggplot(caesarean_ci, aes(x = Caesarean, y = Mean)) +
  geom_point() +
  geom_errorbar(aes(ymin = Trad.lower, ymax = Trad.upper), width = 0.2) +
  ylab("Mean lung capacity")

ggarrange(plt1, plt2, plt3, ncol = 3, labels = "AUTO")
```



c. Produce all the associated tests for assumptions—i.e. the assumptions to be met when deciding whether to use your choice of inferential test or its non-parametric counterpart.

```
two_assum <- function(x) {
  x_var <- var(x)
  x_norm <- as.numeric(shapiro.test(x)[2])
  result <- c(x_var, x_norm)
  return(result)
}

lungs %>%
  group_by(Gender) %>%
  summarise(LungCap_var = round(two_assum(LungCap)[1], 3),
            LungCap_norm = round(two_assum(LungCap)[2], 3))
```

```
# A tibble: 2 x 3
  Gender LungCap_var LungCap_norm
  <chr>        <dbl>        <dbl>
1 female        6.58        0.002
2 male          7.2         0.073
```

```
lungs %>%
  group_by(Smoke) %>%
  summarise(LungCap_var = round(two_assum(LungCap)[1], 3),
            LungCap_norm = round(two_assum(LungCap)[2], 3))
```

```
# A tibble: 2 x 3
  Smoke LungCap_var LungCap_norm
  <chr>       <dbl>        <dbl>
1 no           7.43        0.008
2 yes          3.54        0.622
```

```
lungs %>%
  group_by(Caesarean) %>%
  summarise(LungCap_var = round(two_assum(LungCap)[1], 3),
            LungCap_norm = round(two_assum(LungCap)[2], 3))
```

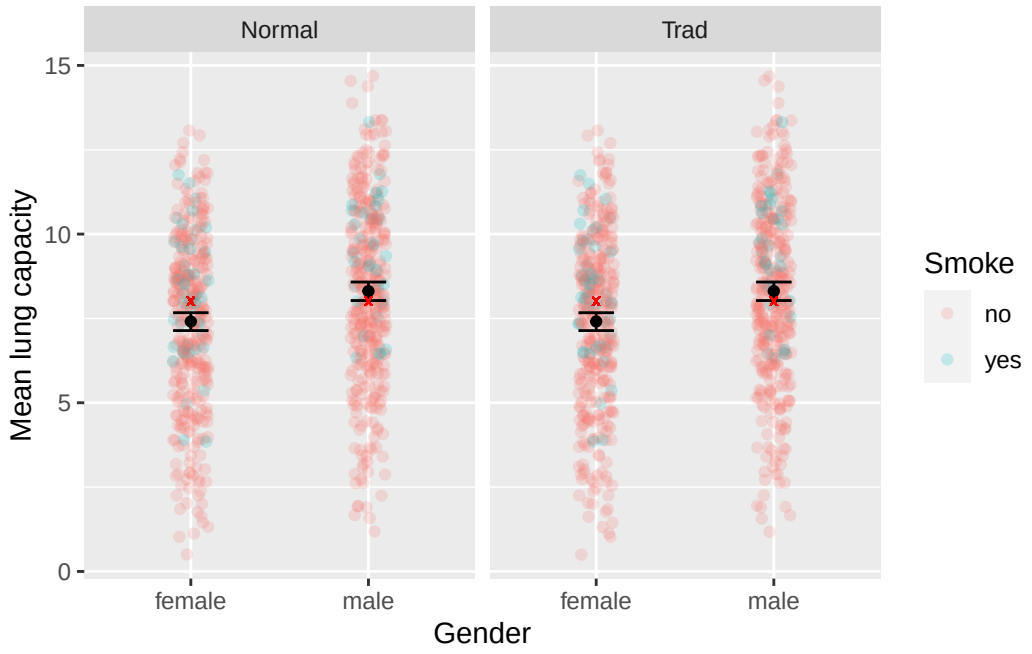
```
# A tibble: 2 x 3
  Caesarean LungCap_var LungCap_norm
```

```
  <chr>           <dbl>       <dbl>
1 no               7.13       0.004
2 yes              6.97       0.554
```

```
# It would be best to continue with a Wilcoxon test
```

d. Create a combined tidy dataframe (observe tidy principles) with the estimates for the 95% CI for the `LungCap` data (`LungCap` as a function of `Gender`), estimated using both the traditional and bootstrapping approaches. Create a plot comprising two panels (one for the traditional estimates, one for the bootstrapped estimates) of the mean, median, scatter of raw data points, and the upper and lower 95% CI.

```
groupwiseMean(LungCap ~ Gender, data = lungs, conf = 0.95, digits = 3, normal = TRUE) |>
  pivot_longer(cols = Trad.lower:Normal.upper,
               names_to = "type", values_to = "CI") |>
  separate(col = type, into = c("type", "direction")) |>
  pivot_wider(names_from = direction, values_from = CI) |>
  ggplot(aes(x = Gender, y = Mean)) +
    geom_jitter(data = lungs, aes(x = Gender, y = LungCap, colour = Smoke),
                width = 0.1, alpha = 0.2) +
    geom_point(colour = "black") +
    geom_errorbar(aes(ymin = lower, ymax = upper),
                  width = 0.2, colour = "black") +
    geom_point(data = lungs, aes(x = Gender, y = median(LungCap)),
               colour = "red", shape = "X") +
    facet_wrap(~type) +
    ylab("Mean lung capacity")
```



e. Undertake a statistical analysis that incorporates both the effect of `Age` *and* one of the categorical variables on `LungCap`. What new insight does this provide?

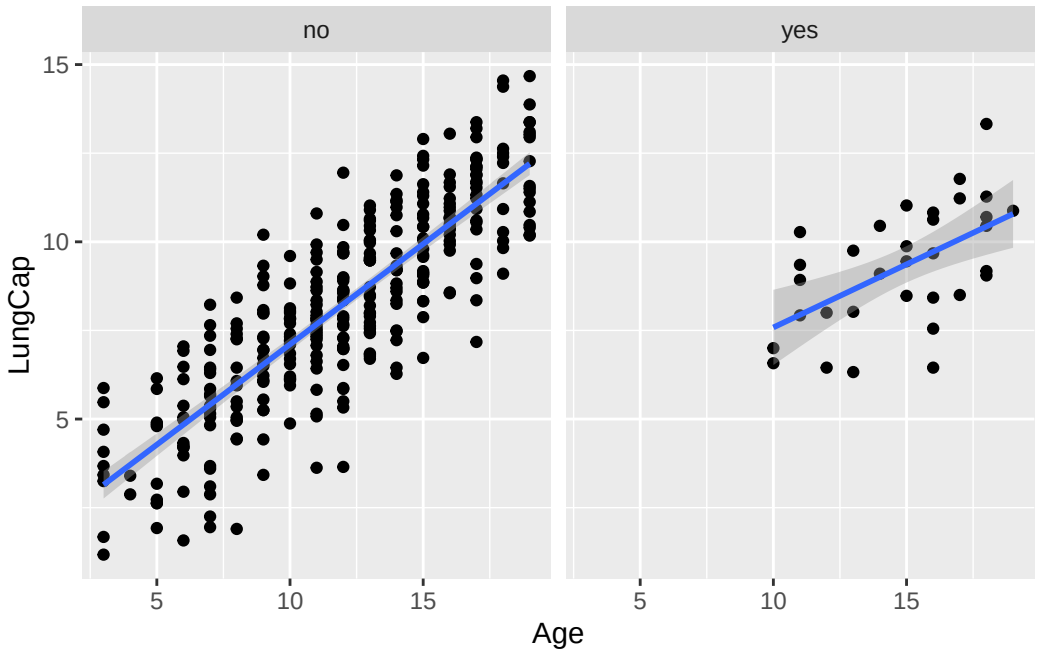
```
# focus only on males
lungs |>
  filter(Gender == "male") |>
  group_by(Smoke) |>
  summarise(LungCap_var = round(two_assum(LungCap)[1], 3),
            LungCap_norm = round(two_assum(LungCap)[2], 3))
```

```
# A tibble: 2 x 3
  Smoke LungCap_var LungCap_norm
  <chr>       <dbl>        <dbl>
1 no           7.52        0.132
2 yes          2.94        0.565
```

```
# above we see that within males, the subgroups based on whether or not
# they smoke are normally distributed in both instances

mod1 <- lm(LungCap ~ Smoke * Age, data = lungs[lungs$Gender == "male", ])
summary(mod1)
```

```
Call:
lm(formula = LungCap ~ Smoke * Age, data = lungs[lungs$Gender ==
    "male", ])

Residuals:
    Min      1Q  Median      3Q     Max
-4.5903 -0.9875  0.0920  1.0286  3.7097

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.44681    0.25191   5.743 1.96e-08 ***
Smokeyes      2.57844    1.48497   1.736   0.0833 .
Age           0.56613    0.01996  28.367  < 2e-16 ***
Smokeyes:Age -0.21021    0.09924  -2.118   0.0348 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.484 on 363 degrees of freedom
Multiple R-squared:  0.6968,    Adjusted R-squared:  0.6943
F-statistic: 278.1 on 3 and 363 DF,  p-value: < 2.2e-16
```

```r
# lung capacity of males is affected by age (disregarding effect of smoke),
# lung capacity is not affected by smoke (disregarding effect of age), but
# there is a significant interaction between them, i.e. the effect of age
# is more pronounced in non-smoker than it is in smokers...

lungs[lungs$Gender == "male", ] |>
  ggplot(aes(x = Age, y = LungCap)) +
    geom_point() +
    geom_smooth(method = "lm") +
    facet_wrap(~Smoke)
```




```
# the figure shows the interaction effect: in non-smokers their lung capacity
# increases more rapidly with age, whereas in smokers, the development of lung
# capacity with age seems to be stunted.
```

0.7 Question 5

0.7.1 The air quality data

Package **datasets**, dataset `airquality`. These are daily air quality measurements in New York, May to September 1973. See the help file for details.

a. Which two of the four response variables are best correlated with each other?

0.8 Question 6

0.8.1 The `shells.csv` data

This dataset contains measurements of shell widths and lengths of the left and right valves of two species of mussels, *Aulacomya* sp. and *Choromytilus* sp. Length and width measurements are presented in mm.

Fully analyse this dataset.

0.9 Question 7

0.9.1 The `fertiliser_crop_data.csv` data

The data represent an experiment designed to test whether or not fertiliser type and the density of planting have an effect on the yield of wheat. The dataset contains the following variables:

- Final yield (kg per acre)—make sure to convert this to the most suitable SI unit before continuing with your analysis
- Type of fertiliser (fertiliser type A, B, or C)
- Planting density (1 = low density, 2 = high density)
- Block in the field (north, east, south, west)

Fully analyse this dataset.

```
fert <- read.csv("../data/fertiliser_crop_data.csv")

# convert to SI units
fert <- fert |>
  mutate(mass = mass / 0.40468564224)

# are assumptions met? note that I also calculate the mean +/- SD here
fert %>%
  group_by(density) %>%
  summarise(mean = mean(mass),
            SD = sd(mass),
            mass_var = round(two_assum(mass)[1], 3),
            mass_norm = round(two_assum(mass)[2], 3))
```

```
# A tibble: 2 x 5
  density   mean    SD mass_var mass_norm
    <int>  <dbl> <dbl>    <dbl>     <dbl>
1       1 11889.  40.8    1668.     0.469
2       2 11920.  43.3    1877.     0.529
```

```
fert %>%
  group_by(block) %>%
  summarise(mean = mean(mass),
            SD = sd(mass),
            mass_var = round(two_assum(mass)[1], 3),
            mass_norm = round(two_assum(mass)[2], 3))
```

```
# A tibble: 4 x 5
  block   mean    SD mass_var mass_norm
  <chr>  <dbl> <dbl>    <dbl>     <dbl>
1 east  11925.  43.4    1882.     0.422
2 north 11894.  42.2    1781.     0.77
3 south 11884.  39.7    1578.     0.212
4 west  11915.  43.7    1906.     0.21
```

```
fert %>%
  group_by(fertilizer) %>%
  summarise(mean = mean(mass),
            SD = sd(mass),
            mass_var = round(two_assum(mass)[1], 3),
            mass_norm = round(two_assum(mass)[2], 3))
```

```
# A tibble: 3 x 5
  fertilizer   mean    SD mass_var mass_norm
  <chr>       <dbl> <dbl>    <dbl>     <dbl>
1 A          11887.  46.1    2122.     0.774
2 B          11899.  38.6    1490.     0.887
3 C          11927.  40.3    1623.     0.254
```

```
# yes, all assumptions check out, proceed with normal paramatric stats

# do an ANOVA and look at main effects first
aov1 <- aov(mass ~ density + block + fertilizer, data = fert)
summary(aov1)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
density      1  23164   23164  15.224 0.000184 ***
block        2   2199    1099   0.723 0.488329
fertilizer   2  27444   13722   9.018 0.000269 ***
Residuals   90 136940    1522
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# the block effect is not significant but density and fertilizer are

# let's check if the fertilizer type interacts with density
```

```r
aov2 <- aov(mass ~ density * fertilizer, data = fert)
summary(aov2)
```

```
                   Df Sum Sq Mean Sq F value   Pr(>F)
density             1  23164   23164  15.195 0.000186 ***
fertilizer          2  27444   13722   9.001 0.000273 ***
density:fertilizer  2   1935     967   0.635 0.532500
Residuals          90 137203    1524
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# no interaction effect is present, so the fertilizer has the same
# effect regardless of at which planting density it is applied

# lets see which planting fertilizer results in the greatest mass

TukeyHSD(aov2, which = "fertilizer", ordered = TRUE)
```

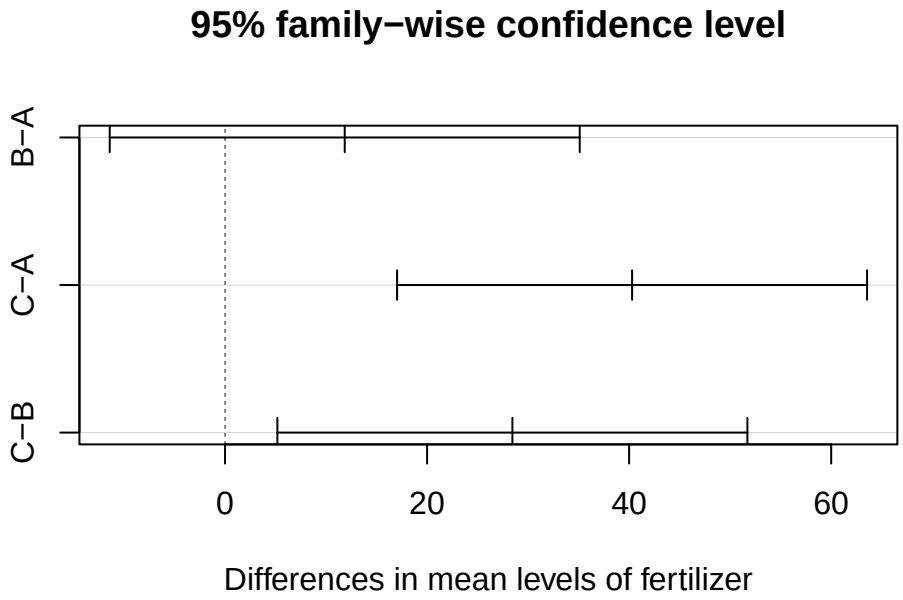
```
  Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = mass ~ density * fertilizer, data = fert)

$fertilizer
        diff        lwr      upr     p adj
B-A 11.84752 -11.414312 35.10935 0.4482026
C-A 40.29177  17.029945 63.55360 0.0002393
C-B 28.44426   5.182428 51.70609 0.0123951
```

```r
TukeyHSD(aov2, which = "fertilizer", ordered = TRUE)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = mass ~ density * fertilizer, data = fert)

$fertilizer
```

```
        diff        lwr      upr     p adj
B-A 11.84752 -11.414312 35.10935 0.4482026
C-A 40.29177  17.029945 63.55360 0.0002393
C-B 28.44426   5.182428 51.70609 0.0123951
```

```
plot(TukeyHSD(aov2, which = "fertilizer", ordered = TRUE))
```




```
# here we can see that the mass of crop produced by fertilizer C is the
# greatest, significantly more so compared to both A and B; the effect
# of fertilizer B is no different than that of A
#
# the second planting density also yields a greater mass per ha
#
# make sure the results are written up as appropriate for a journal,
# so indicate the d.f., S.S., and p-value
```

0.10 The end

Submit the .html file wherein you provide answers to Questions 1–7 by no later than 19:00 today. Label the script as follows:

BCB744_<Name>_<Surname>_Summative_Task_2.html, e.g.

BCB744_AJ_Smit_Summative_Task_2.html.

Upload your .html files onto Google Forms.