CWRU DSCI351-351m-451: Lab Exercise LE5 NAME

Inference-t-tests-CIs

Prof.:Roger French, Paul Leu TA: Sameera Nalin Venkat, Raymond Wieser, Mingxuan Li

04 November, 2021

Contents

5.0.1 LE5, 10 points, 2 questions.

5.0.1.1 Lab Exercise (LE) 5

- The purpose of this assignment is to show
  - the amount of improvement students have made
    - in their ability to make and write reproducible code.

The assignment is purposefully simple

- to allow students time to focus on
  - perfecting their report presentation.

THE MAJORITY OF THE GRADE OF THIS ASSIGNMENT WILL BE ASSIGNED

- BASED ON THE ADHERENCE TO PROPER CODE STYLING
  - AS DISCUSSED DURING CLASS AND IN
  - `2-class/2108-351-351m-451-w08a-p1-ClassAndCodingExpectations.pdf`
- And making high quality plots and figures
  - That communicate clearly and coherently
- And having appropriate

  - Comments in code blocks
  - And discussion in the Rmarkdown sections

Questions 1 and 2 use beaver body temperature data

- from the `beavers` datasets
  - built into base R.

Question 3 uses raw data to conduct a brief EDA

Also in the Links section at the end of LE5

- are some useful references to discussion of t-tests and CIs
- which can be useful

5.1 LE5.1. Comparisons with one sample t-test (3 points)

Let's start by comparing beaver1's temperature to human body temperature.

We will consider two scenarios in which to conduct our one sample t-tests:

- the first where we pretend to know the variance of the population
  - of all *Castor Canadensis* in Wisconsin in 1994
- and a second, more realistic scenario where we consider
  - the variance of the population to be unknown.

We will also vary the situations by using

- a two sided test for scenario 1,
  - where we simply test if the average temperature
  - is significantly different compared to average human body temperature
  - (use 37 ^{o}C),
- and a one sided test for scenario 2,
  - where we test if it is significantly lower
  - than that of an average dog's body temperature
  - (use 38.5 ^{o}C).

We will use 95% confidence interval

- a 5% significance level.

Important:

- Scenario 1
  - two sided hypothesis: different than human body temperature (37 ^{o}C)
  - known population variance $\sigma^2 = 1$
- Scenario 2: one sided hypothesis, unknown population variance
  - one sided hypothesis: less than dog body temperature (38.5 ^{o}C)
  - unknown population variance
- We are satisfied with 95% confidence

5.1.1 LE5.1.1 (0.5 points)

Data cleaning:

Create a list from the temp variable in the beaver1 dataset. %%% done

```
raw_beaver1 <- beaver1
temp <- raw_beaver1$temp
```

```
df_as_list <- list(raw_beaver1$temp)    # Convert one column to list element
class(df_as_list)
```

```
## [1] "list"
```

```
## Median : 36.87
## Mean : 36.86
```

5.1.2 LE5.1.2 (0.5 point)

What assumptions do we need to make

- before conducting our statistical tests?

ANSWER=>
- The data, collected from a representative and randomly selected portion of the total population, should be independent of one another.

- The dependent variable here, temperature must be measured on a continuous or ordinal scale.
- The observations should follow a normal distribution, if the sample size < 30. If the sample size is large, this constraint can be lifted

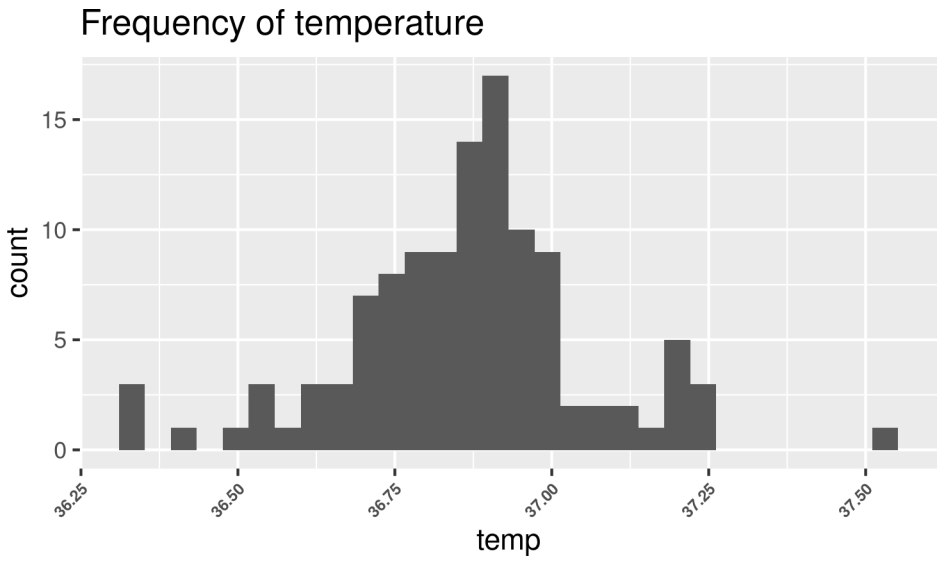
---

Represent visually that the data meets our assumption of normality.

- (hint: histogram)

```
t <-  theme(legend.text = element_text(size = 4) ) +
theme(legend.key.size = unit(0.5, 'cm') ,
axis.text.x = element_text(size = 6,
face = "bold", angle = 45, hjust = 1))

ggplot(data = raw_beaver1, aes(x = temp)) +
  geom_histogram() + labs(title = "Frequency of temperature") + t
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
## As we can see the temperature variable follows a normal distribution,
## with a mean between (36.75 and 37)
```

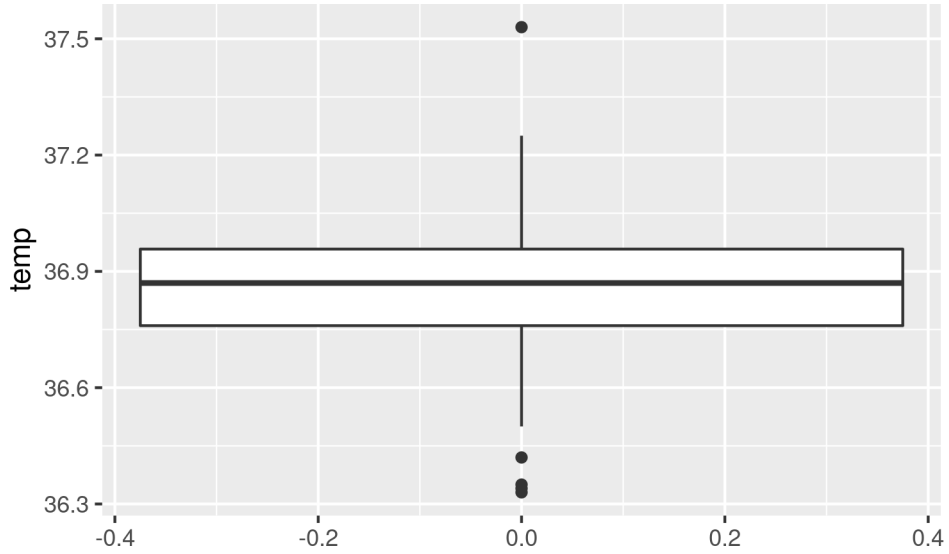
5.1.3 LE5.1.3 (1 points)

Scenario 1: Variance of the population is known $\sigma^2 = 1$,

- testing if different from human body temp
  - (use 37 ^{o}C)

Visualize the sample using a boxplot.

```
ggplot(data = raw_beaver1, aes(y = temp)) +
  geom_boxplot()
```




```
## As we can see the temperature variable follows a normal distribution,
## with a median  less than 36.9
```

Follow the four steps of hypothesis testing:

- State the hypothesis
- Calculate the test statistic
- Calculate the critical value
- Draw your conclusion

```
## Method 1 - since variance of the population is known we use normal
## distribution

mean_temp_beaver_sample = mean(raw_beaver1$temp)
mean_temp_human_body = 37 ## given in question
n =  length(raw_beaver1$temp)
variance = 1 ## given in question
significance_level = 5/100

std_deviation = sqrt(variance)
point_estimate = mean_temp_beaver_sample
null_value = mean_temp_human_body
standard_error = std_deviation / sqrt(n)
```

```
z_score = (point_estimate - null_value)/standard_error
p_value = 2 * pnorm(z_score, lower.tail = TRUE) ## 2- sided
# In an upper-tailed test the decision rule has investigators reject H0 if the
# test statistic is larger than the critical value. In a lower-tailed test the
# decision rule has investigators reject H0 if the test

if(p_value > significance_level){
  glue("The chance of occurence of  {point_estimate} is greater than
        {significance_level} = > there is something going on and we reject H0")
}else
    glue("The chance of occurence of {point_estimate} is lower than
         {significance_level} = > point estimate came due to sampling
         variability. There is nothing going on and we reject H1")
```

```
## The chance of occurence of  36.8621929824561 is greater than
## 0.05 = > there is something going on and we reject H0
```

```
critical_val_upper = point_estimate + 1.96 * standard_error
critical_val_lower = point_estimate - 1.96 * standard_error

glue(" The confidence intervel lies in between {critical_val_lower} and  {critical_val_upper}")
```

```
##  The confidence intervel lies in between 36.6786221633861 and  37.0457638015262
```

ANSWER=>

- Looking at the above two plots it seems that beaver body temp is quite close to human temp, but to be certain with 95 % confidence level and 5% significance level, we must undergo the hypothesis testing
- H0 : Mean_temp_beaver == Human body temp (37 c)
- H1 : Mean_temp_beaver != Human body temp (37 c)
- The chance of occurrence of 36.8621929824561 is greater than 0.05 = > there is something going on and we reject H0
- The confidence intervel lies in between 36.6786221633861 and 37.0457638015262

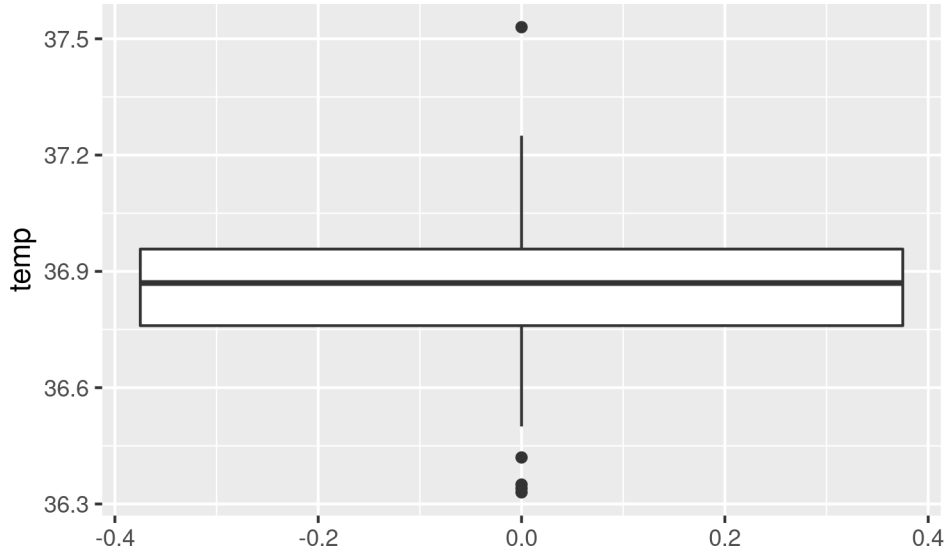
5.1.4 LE5.1.4 (1 points)

Scenario 2: Variance of the population is unknown,

- testing if less than dog body temp
  - (use 38.5 ^{o}C)

Visualize the sample using a boxplot.

```
ggplot(data = raw_beaver1, aes(y = temp)) +
  geom_boxplot()
```



Follow the steps of hypothesis testing.

What changes about our calculations when the variance is unknown?

- State the hypothesis
- Calculate the test statistic
- Calculate the critical value
- Draw your conclusion

```
##Method 2, since we do not know the true variance of the population
## We use t test.

test <- t.test( raw_beaver1$temp,
  mu = mean_temp_dog_body
  alternative = "less"
)
```

```
## Error in t.test.default(raw_beaver1$temp, mu = mean_temp_dog_body, alternative = "less"): object 'mea
```

```
test
```

```
## Error in eval(expr, envir, enclos): object 'test' not found
```

```
## unused code

# std_deviation = sd(raw_beaver1$temp) # since standard deviation is not given of
# # the beaver population, we build standard deviation from the given sample
# mean_temp_beaver_sample = mean(raw_beaver1$temp)
# mean_temp_dog_body  =  38.5 ## given in question
# n =  length(raw_beaver1$temp)
# significance_level = 5/100
#
# point_estimate = mean_temp_beaver_sample
# null_value = mean_temp_dog_body
# standard_error = std_deviation / sqrt(n)



# z_score = (point_estimate - null_value)/standard_error
# p_value = pnorm(z_score) ## 1- sided
```

```
#
# if(p_value > significance_level){
#   glue(" The chance of occurence of {point_estimate} is greater than
#         {significance_level} = > there is something going on and we reject H0 ")
# }else
#     glue(" The chance of occurence of {point_estimate} is lower than
#          {significance_level} = > point estimate came due to sampling
#          variability. There is nothing going on and we reject H1 ")
#
# critical_val_upper = point_estimate + 1.96 * standard_error
# critical_val_lower = point_estimate - 1.96 * standard_error
```

ANSWER =>

- H0 : The body temp of beaver is close to mean dog body temp : 38.5
- H1 : The body temp of beaver is significantly lower than that of dog temp.
- Since the p-value : 1.536906e-107 which is very less than 0.05, therefore we can reject the null hypothesis for the alternate that : The body temp of beaver is
  significantly lower than that of dog temp.
- 95 percent confidence interval, that the true mean beaver temp lies in between: -Inf 36.89224

5.2 LE5.2. Comparing two samples with Student's t-test (5 points)

Next, we'll compare beaver1 and beaver2.

We'll use a two sided test

- under the realistic conditions of two independent samples
  - with equal but unknown population variances.

Our hypothesis is that the beavers

- have the same average temperature.

Important:

- Independent samples
- Equal, unknown variances
- Testing that beavers have the same average temperature $\mu_1 - \mu_2 = 0$

5.2.1 LE5.2.1 (0.5 points)

Data cleaning:

To compare similar datasets,

- let's use the same timeframe for both beavers.
- Use the datapoints between 9:30 and 23:50, inclusive,
  - and create lists for each beaver's temperature.

```
## time range [0930 to 2350]
## subset from both dataset and select only temp

temp_beaver1 <- beaver1 %>%
  filter(time >= 0930 &  time <= 2350)%>%
  select(temp)
```

```
list_temp_beaver1 <- list(temp_beaver1$temp)

temp_beaver2 <- beaver2 %>%
  filter(time >= 0930 &  time <= 2350)%>%
  select(temp)


list_temp_beaver2 <- list(temp_beaver2$temp)
```

5.2.2 LE5.2.2 (1 point)

What other possible scenarios exist

- with two samples for the Student's t-test?

ANSWER=> There are 5 scenarios:

1. Two independent samples

- 2 variances are known
- 2 variances are unknown, but equal
- 2 variances are unknown, and unequal

2. Two paired samples (Which means they are from the same experiment sample)

- Variances of the difference is known
- Variances of the difference is unknown

What assumptions do we have to make

- before conducting our statistical test ?

ANSWER =>

Variable type: - A Student's t-test requires a mix of one quantitative dependent variable (which corresponds to the measurements to which the question relates) and one qualitative independent variable (with exactly 2 levels which will determine the groups to compare). Here, we have temperature as quantitative dependent variable which is dependent on the type of beaver population.

Independence: - The data, collected from a representative and randomly selected portion of the total population, should be independent between groups and within each group. Here, Beaver 1 and Beaver 2 data is independent of each other.

Normality : - With samples (n $<$ 30), normality of both samples is required. With large samples (n 30), normality of the data is not required. By the central limit theorem, sample means of large samples are often well-approximated by a normal distribution even if the data are not normally distributed. It is therefore not required to test the normality assumption when the number of observations in each group/sample is large.

Equality of variances: - When the two samples are independent, the variances of the two groups should be equal in the populations.

Outliers: - Outlier is a value or an observation that is distant from the other observations. There should be no significant outliers in the two groups.

5.2.3 LE5.2.3 (1 point)

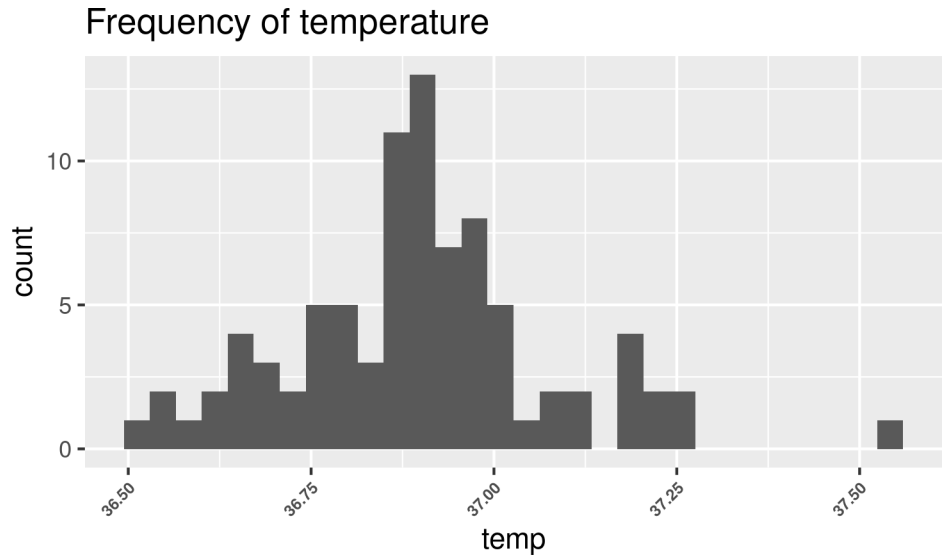
What is the Central Limit Theorem?

When we collect a sufficiently large sample of n independent observations from a population with mean mu and standard deviation σ, the sampling distribution of x_bar will be nearly normal with Mean = mu and Standard Error = σ / sqrt(n)

Check the normality of the lists of beaver temperature.

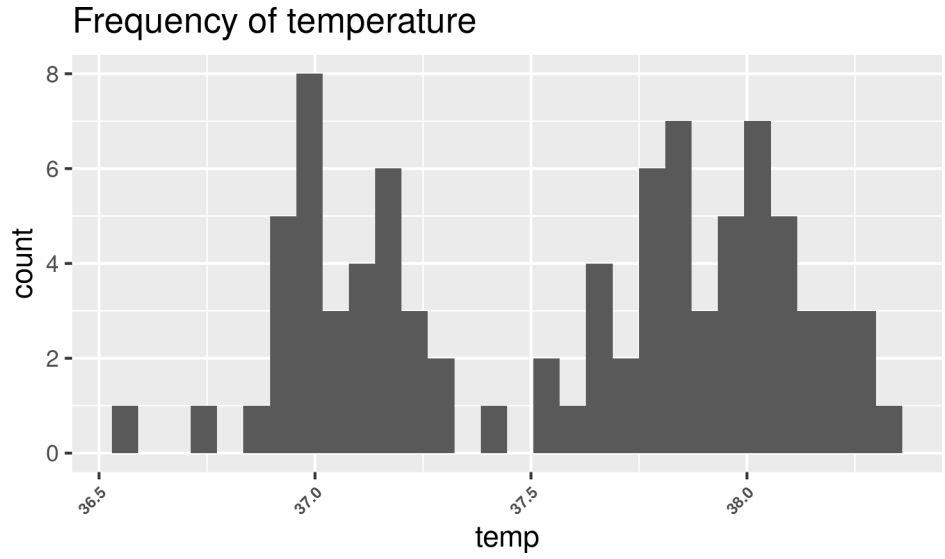
```
ggplot(data = temp_beaver1, aes(x = temp)) +
  geom_histogram() + labs(title = "Frequency of temperature") + t
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## somewhat normally distributed

ggplot(data = temp_beaver2, aes(x = temp)) +
  geom_histogram() + labs(title = "Frequency of temperature") + t
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
Mean_beaver1   :36.90
```

```
## Error in eval(expr, envir, enclos): object 'Mean_beaver1' not found
```

```
mean_beaver2 : 37.58
```

```
## Error in eval(expr, envir, enclos): object 'mean_beaver2' not found
```

```
## not at all normally distributed, infact it seems to have two peaks, and a
## clear dip in the center
```

Do we need the data to be normal

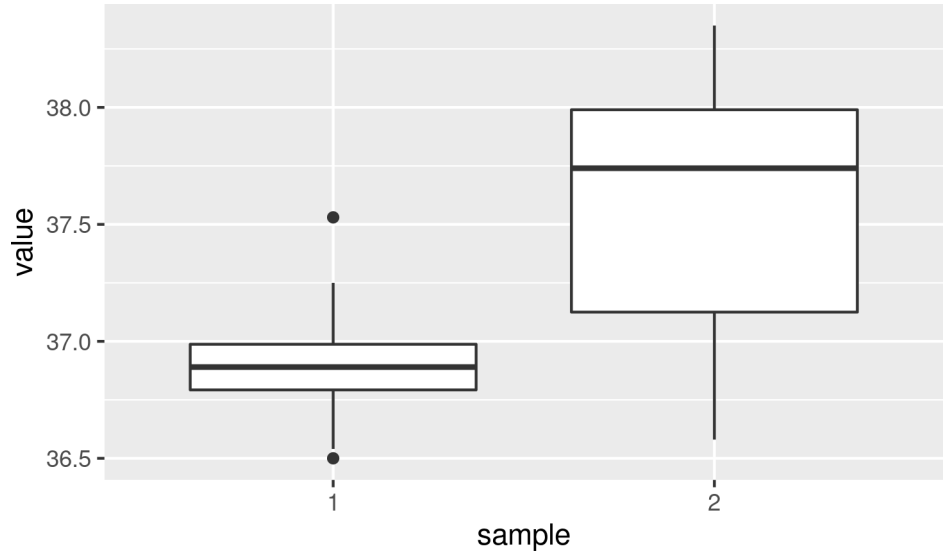
- to carry out our test?
- Why or why not?

ANSWER=> The data does not have to be normal in either samples, because samples have a size greater than 30, since we are doing a t-test, we are using t-distribution that accounts for this variability by using the sample variance accordingly.

5.2.4 LE5.2.4 (1 point)

Compare the two samples using boxplots.

```
## to visualize both the  samples on the same  plot
combo <- bind_rows(temp_beaver1, temp_beaver2)

dat_ggplot <- data.frame(
  value = combo$temp,
  sample = c(rep("1", 86), rep("2", 87))
)


ggplot(data = dat_ggplot, aes(y = value, x = sample)) +
  geom_boxplot()
```




```
summary(temp_beaver1)
```

```
##       temp
```

```
##  Min.   :36.50
##  1st Qu.:36.79
##  Median :36.89
##  Mean   :36.90
##  3rd Qu.:36.99
##  Max.   :37.53
```

```
summary(temp_beaver2)
```

```
##       temp
##  Min.   :36.58
##  1st Qu.:37.12
##  Median :37.74
##  Mean   :37.58
##  3rd Qu.:37.99
##  Max.   :38.35
```

```
## can you have a side by side box plot
## They have to be part of the same dataset


### beaver 1 - median 36.89
### beaver 2 - median 37.74
### on the surface it seems that both of them are very different.
```

5.2.5 LE5.2.5 (1.5 points)

Independent samples with equal, unknown variances.

- Hypothesis is that the beavers have the same average temperature.

Follow the steps of hypothesis testing.

What changes about our calculations when the variance is unknown?

- State the hypothesis
- Calculate the test statistic
- Calculate the critical value
- Draw your conclusion

```
## This is an unpaired test, because both the samples are independent of each other
## Since variance is unknown we must find it from the samples



test <- t.test(temp_beaver1$temp, temp_beaver2$temp,
  var.equal = TRUE, alternative = "two.sided"
)
test
```

```
##
##  Two Sample t-test
##
## data:  temp_beaver1$temp and temp_beaver2$temp
## t = -12.588, df = 171, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -0.7890586 -0.5751435
## sample estimates:
## mean of x mean of y
##  36.89721  37.57931
```

```
test$p.value
```

```
## [1] 3.884138e-26
```

```
# unused code
# ## Beaver1
# mean_temp_beaver1 = mean(temp_beaver1$temp)
# n1 = length(temp_beaver1$temp)
# std_deviation1 = sd(temp_beaver1$temp)
#
# ## Beaver2
# mean_temp_beaver2 = mean(temp_beaver2$temp)
# n2 = length(temp_beaver2$temp)
# std_deviation2 = sd(temp_beaver2$temp)
#
#
# point_estimate = mean_temp_beaver1 -  mean_temp_beaver2
# null_value = 0
# standard_error = sqrt((std_deviation1^2/n1 ) + (std_deviation2^2/n2 ))

# if(p_value > significance_level){
#   glue(" The chance of occurence of {point_estimate} is greater than
#          {significance_level} = > there is something going on and we reject H0 ")
# }else
#     glue(" The chance of occurence of {point_estimate} is lower than
#          {significance_level} = > point estimate came due to sampling
#          variability. There is nothing going on and we reject H1 ")
#
# critical_val_upper = point_estimate + 1.96 * standard_error
# critical_val_lower = point_estimate - 1.96 * standard_error
```

ANSWERS ->

- H0 : The mean temp of both of the population of beavers is the same. mu1 - mu2 = 0
- H1 : The mean temp of both of the population of beavers is not the same. mu1 - mu2 != 0
- Test Statistic : p_value =3.884138e-26 < 0.05 Since the p_value is very less than the significant value, we reject the null hypothesis for the alternate, and say that the mean of both population is not the same.
- With 95 % Confidence level , the true mean lies between (-0.7890586 -0.5751435)

5.3 LE5.3. Data Cleaning of Palmer's Penguins (2 points)

We will now perform some elementary data cleaning and EDA.

The Palmer's Penguins dataset includes measurements of penguin characteristics used in sampling population dynamics.

Included in the palmerspenguins data set are - Raw Data File - Cleaned Data File

Our goal is to perform some elementary operations on the RAW data in order to use the cleaned data for a future time.

5.3.1 LE5.3.1 (.25 points)

Raw data files often contain improperly named columns that are not in accordance with R naming conventions.

Two commonly used naming conventions are camelCase and snake_case

Using one of the above naming conventions

```
- Rename the columns
    - Avoiding Special Characters
    - Spaces
    - "dashes"
```

```
library(palmerpenguins)

??palmerpenguins
```

```
## No vignettes or demos or help files found with alias or concept or
## title matching 'palmerpenguins' using fuzzy matching.
```

```
df <- palmerpenguins::penguins_raw


# rename code : titanic_df <- titanic_df %>%
# rename(pc_class = PC)
## Names of dimensioned variables and constants
##  should usually have a units suffix.

## we will be using snake_case for columns
df <- df %>%
  rename(  study_name = studyName, sample_number = `Sample Number`,
           species = Species, region = Region,
           island = Island, stage = Stage,
           individual_id = `Individual ID`,
           clutch_completion = `Clutch Completion`, date_egg = `Date Egg`,
           culmen_length_mm = `Culmen Length (mm)`,
           culmen_depth_mm = `Culmen Depth (mm)`,
           flipper_length_mm = `Flipper Length (mm)`,
           body_mass_g = `Body Mass (g)`, sex = Sex,
           delta_15_n_o_oo = `Delta 15 N (o/oo)`,
           delta_13_c_o_oo = `Delta 13 C (o/oo)`,
           comments = Comments   )

colnames(df)
```

```
##  [1] "study_name"        "sample_number"     "species"
##  [4] "region"            "island"            "stage"
##  [7] "individual_id"     "clutch_completion" "date_egg"
## [10] "culmen_length_mm"  "culmen_depth_mm"   "flipper_length_mm"
## [13] "body_mass_g"       "sex"               "delta_15_n_o_oo"
## [16] "delta_13_c_o_oo"   "comments"
```

5.3.2 LE5.3.2 (.5 points)

Raw data files often contain excess data that will not be used in our analysis. Sometimes extraneous data is contained in cells with important data.

Let's try to seperate out some of the useful information from the less usefull

```
- Seperate the "Species" column into one for
    - Scientific Name
    - Common Name

- Seperate the "Date Egg" column into one for
    - Day
    - Month
    - Year
```

```
# code : myfile %>% mutate(V5 = ifelse(V1 == 1 & V2 != 4, 1,
# ifelse(V2 == 4 & V3 != 1, 2, 0)))
# code : my_basket %>% mutate(Price_band = case_when(Price>=50 &
# Price <=70   ~ "Medium", Price > 70 ~ ##"High", TRUE ~ "Low"))

df <- df %>% mutate(
  common_name =
    case_when(species =="Adelie Penguin (Pygoscelis adeliae)"   ~ "Adelie Penguin",
              species =="Gentoo penguin (Pygoscelis papua)"   ~ "Gentoo penguin",
            species =="Chinstrap penguin (Pygoscelis antarctica)"   ~ "Chinstrap penguin"))

df <- df %>% mutate(
  scientific_name =
    case_when(species =="Adelie Penguin (Pygoscelis adeliae)"   ~ "Pygoscelis adeliae",
              species =="Gentoo penguin (Pygoscelis papua)"   ~ "Pygoscelis papua",
            species =="Chinstrap penguin (Pygoscelis antarctica)"   ~ "Pygoscelis antarctica"))

df <- select(df, -species)



# code : df %>% separate(x, c("key", "value"), ": ", extra = "merge")
df <- df %>% separate(date_egg, c("year", "month", "day"), "-", extra = "merge")



glimpse(df)
```

```
## Rows: 344
## Columns: 20
## $ study_name        <chr> "PAL0708", "PAL0708", "PAL0708", "PAL0708", "PAL0708~
## $ sample_number     <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1~
## $ region            <chr> "Anvers", "Anvers", "Anvers", "Anvers", "Anvers", "A~
## $ island            <chr> "Torgersen", "Torgersen", "Torgersen", "Torgersen", ~
## $ stage             <chr> "Adult, 1 Egg Stage", "Adult, 1 Egg Stage", "Adult, ~
## $ individual_id     <chr> "N1A1", "N1A2", "N2A1", "N2A2", "N3A1", "N3A2", "N4A~
## $ clutch_completion <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "No", "No"~
## $ year              <chr> "2007", "2007", "2007", "2007", "2007", "2007", "200~
## $ month             <chr> "11", "11", "11", "11", "11", "11", "11", "11", "11"~
```

```
## $ day               <chr> "11", "11", "16", "16", "16", "16", "15", "15", "09"~
## $ culmen_length_mm  <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, ~
## $ culmen_depth_mm   <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18.1, ~
## $ flipper_length_mm <dbl> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, 186~
## $ body_mass_g       <dbl> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 3475, ~
## $ sex               <chr> "MALE", "FEMALE", "FEMALE", NA, "FEMALE", "MALE", "F~
## $ delta_15_n_o_oo   <dbl> NA, 8.94956, 8.36821, NA, 8.76651, 8.66496, 9.18718,~
## $ delta_13_c_o_oo   <dbl> NA, -24.69454, -25.33302, NA, -25.32426, -25.29805, ~
## $ comments          <chr> "Not enough blood for isotopes.", NA, NA, "Adult not~
## $ common_name       <chr> "Adelie Penguin", "Adelie Penguin", "Adelie Penguin"~
## $ scientific_name   <chr> "Pygoscelis adeliae", "Pygoscelis adeliae", "Pygosce~
```

5.3.3 LE5.3.3 (.25 points)

Additionally, sometimes data will contain variables that you dont recognize.

- What is Delta 15 N?

ANSWER -> a number denoting the measure of the ratio of stable isotopes 15N:14N

- What is Delta 13 C?

ANSWER -> a number denoting the measure of the ratio of stable isotopes 13C:12C

- What is a penguins culmen?

ANSWER -> The dorsal ridge of a bird's bill

Rename the "culmen" variable into plain English using tidyverse functions.

```
df <- df %>%
  rename(dorsal_ridge_of_bill_length_mm = culmen_length_mm,
         dorsal_ridge_of_bill_depth_mm = culmen_depth_mm)
```

5.3.4 LE5.3.4 (1 points)

Finally, the raw data file has some additional columns of information

- What is Delta 15 N?
- What is Delta 13 C?

ANSWER -> - What is Delta 15 N? - a number denoting the measure of the ratio of stable isotopes 15N:14N - What is Delta 13 C? - a number denoting the measure of the ratio of stable isotopes 13C:12C

Develop a DATA SCIENCE QUESTION that you could ask using the Delta15N and Delta13C data.

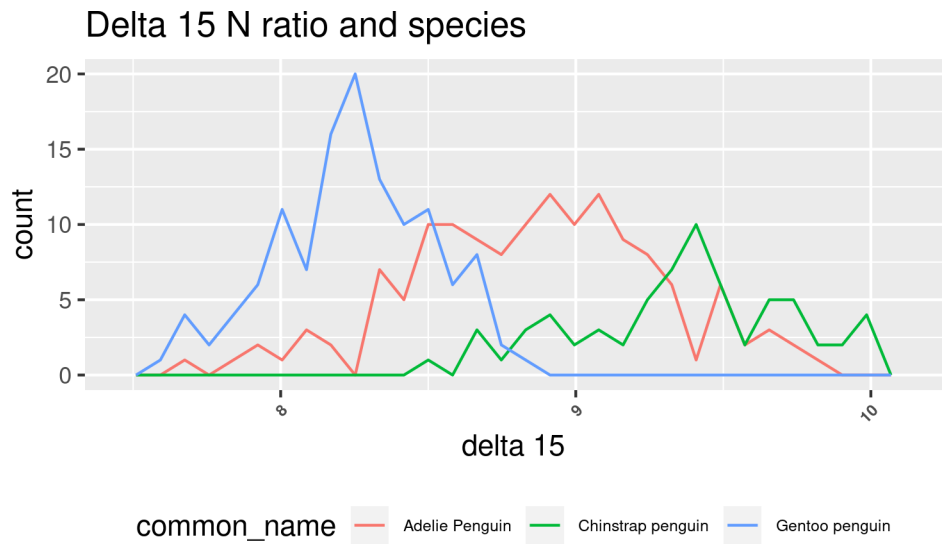
ANSWER -> The ratio between the heavy, stable isotope of carbon and the normal isotope in a sample of interest. Since organisms take up C12 in preference to C13, the ratio is used to determine whether or not the carbon in the specimen is of biological origin.

- Do penguins of same species have same values of delta N and delta C ?

Using ggplot2 visualize the question you are trying to answer

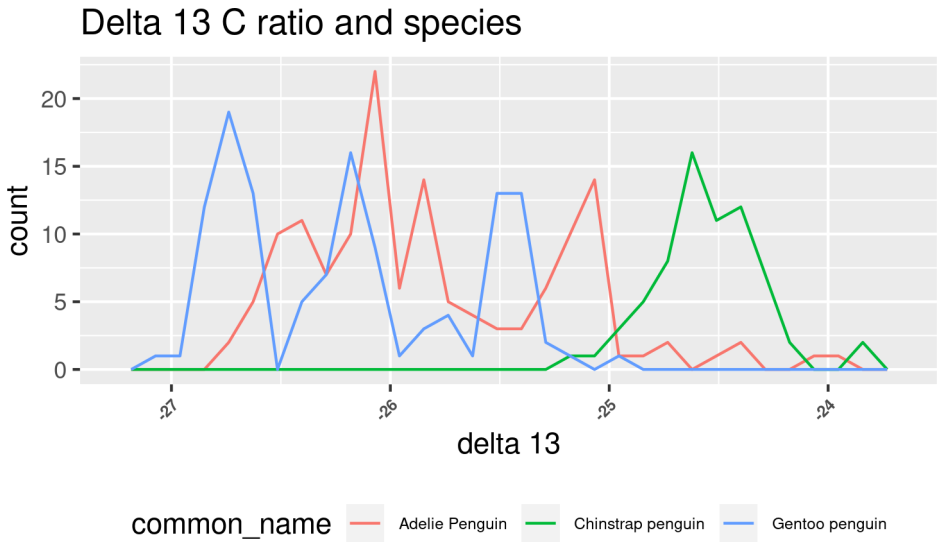
```
 t <-  theme(legend.text = element_text(size = 6)) +
theme(legend.key.size = unit(0.5, 'cm') ,
axis.text.x = element_text(size = 6,
face = "bold", angle = 45, hjust = 1))

ggplot(df, aes(x = delta_15_n_o_oo, color = common_name)) +
geom_freqpoly(bin_width = 20) +
labs(title = "Delta 15 N ratio and species", x = " delta 15 ")+
t + theme(legend.position = "bottom")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(df, aes(x = delta_13_c_o_oo, color = common_name)) +
geom_freqpoly(bin_width = 20) +
labs(title = "Delta 13 C ratio and species", x = " delta 13 " )+
 t + theme(legend.position = "bottom")
```

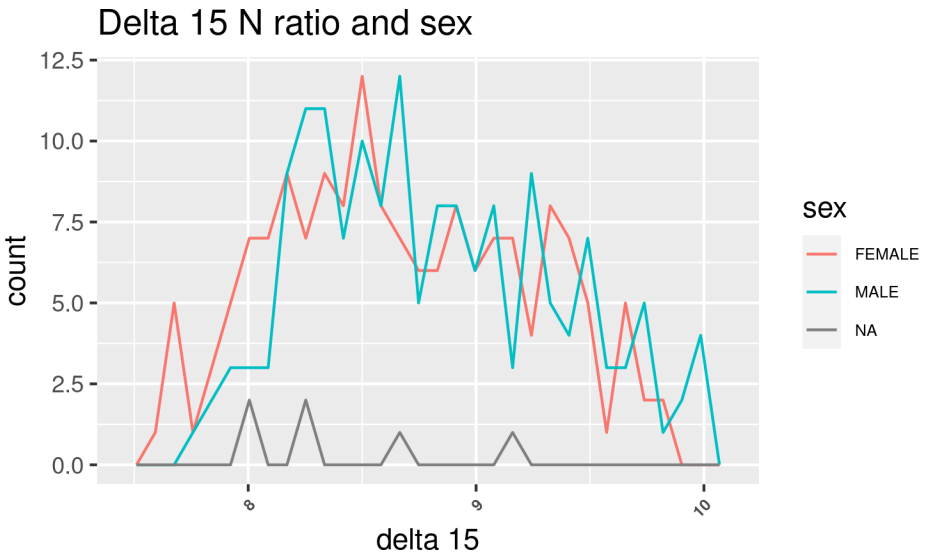
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## this tells us, that penguins of a particular species have the same delta values
## atleast they seem to cluster around one side.
```

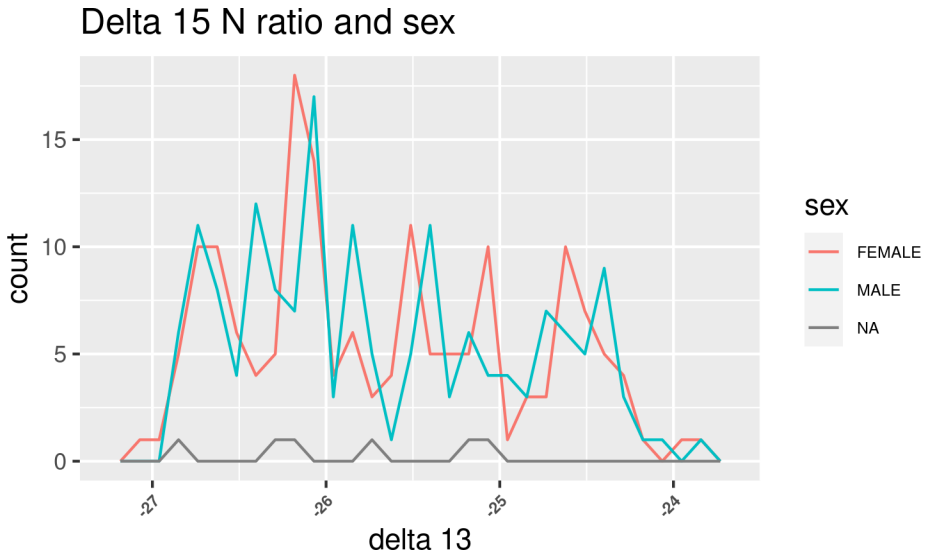
```
ggplot(df, aes(x = delta_15_n_o_oo, color = sex)) +
geom_freqpoly(bin_width = 20) +
labs(title = "Delta 15 N ratio and sex" , x = " delta 15 ")  +
t
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(df, aes(x = delta_13_c_o_oo, color = sex)) +
geom_freqpoly(bin_width = 20) +
labs(title = "Delta 15 N ratio and sex", x = " delta 13 " )+
t
```

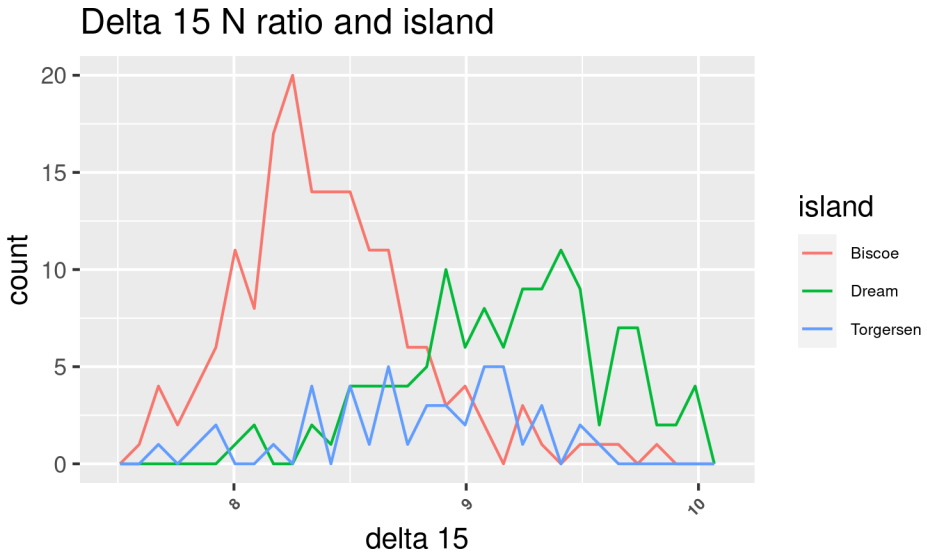
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## both sexes are found in both the columns uniformly, there can be no distinction
## based on sex
```

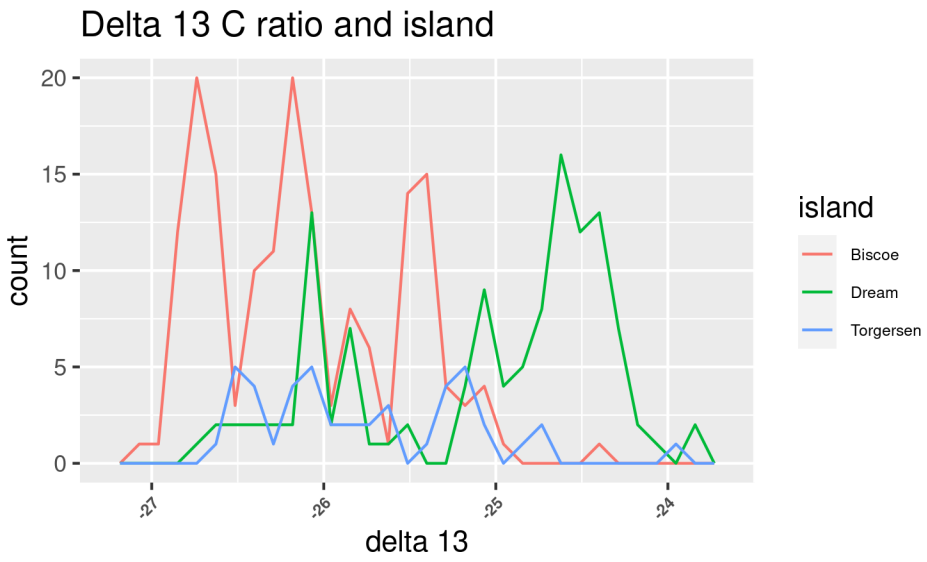
```
ggplot(df, aes(x = delta_15_n_o_oo, color = island)) +
geom_freqpoly(bin_width = 20) +
labs(title = "Delta 15 N ratio and island", x = " delta 15 " ) + t
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(df, aes(x = delta_13_c_o_oo, color = island)) +
geom_freqpoly(bin_width = 20) +
labs(title = "Delta 13 C ratio and island", x = " delta 13 " ) + t
```

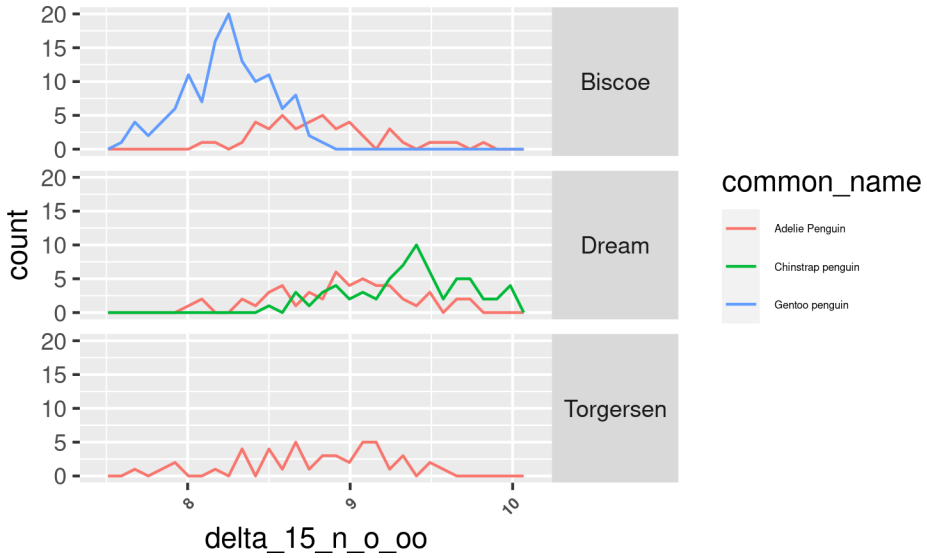
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## there seems to be a relationship between delta levels and island
## most penguins from Biscoe seem to have delta 15 values between (7.5 -9)
## from dream (9 - 10)
## from Torgersen (8.25 - 9.25)

## delta 13 C ratios also reflect the smaller values for Biscoe, the larger
## values for dream and the central values for Torgersen
```

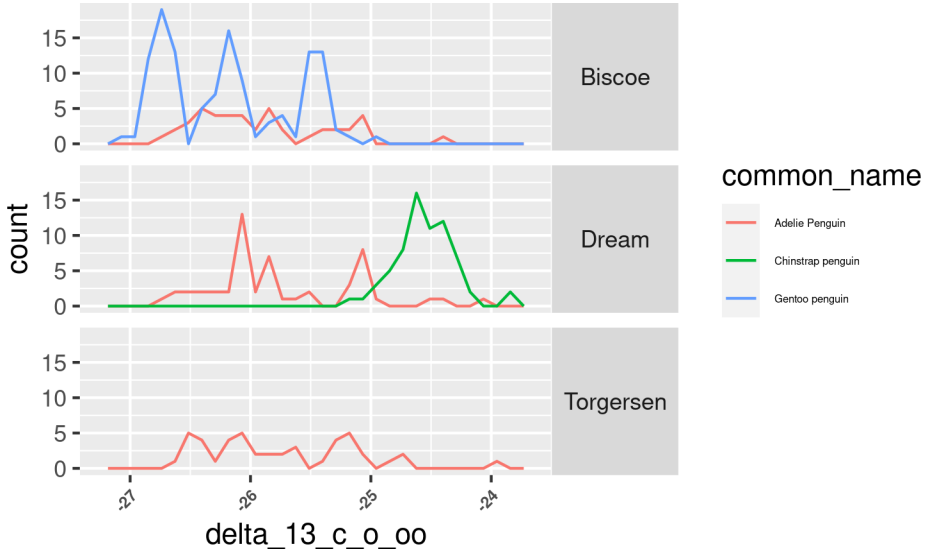
```
## let us put these infos together

 t <-  theme(legend.text = element_text(size = 4) ) +
theme(legend.key.size = unit(0.5, 'cm') ,
axis.text.x = element_text(size = 6,
face = "bold", angle = 45, hjust = 1)) +
  theme(strip.text.y = element_text(angle = 0))

ggplot(df,
aes(x = delta_15_n_o_oo, color = common_name)) + geom_freqpoly(bin_width = 20) +
facet_grid(island ~ .) + t
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(df,
aes(x = delta_13_c_o_oo, color = common_name)) + geom_freqpoly(bin_width = 20) +
facet_grid(island ~ .) + t
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

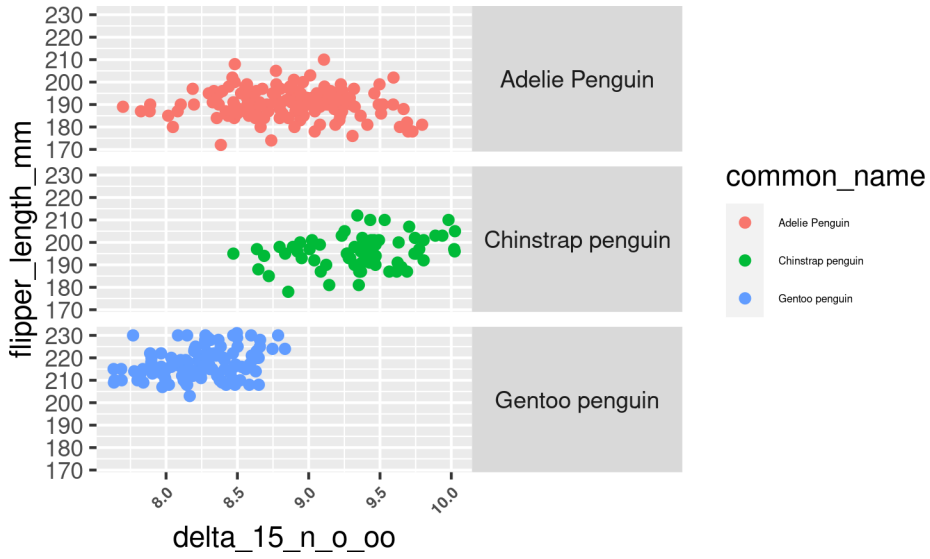



```
# in the island :
#   biscoe, you can only find Adelie and Gentoo penguins and
#   gentoo penguins have the smaller delta 15 values
#
#   Dream, you can only find Adelie and Chinstrap penguins
#   and chin strap penguins have larger  delta 15 values
#
#   Torgersen , you can only find Adelie penguins, they have the central
#   delta 15 values

## delta 13 value has the same pattern
```

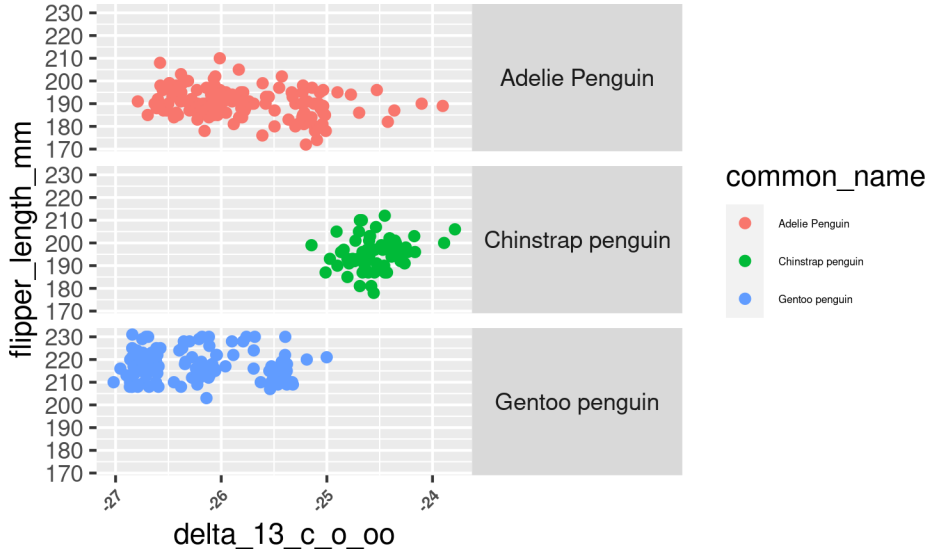
- Is their a relationship between physical features and delta values ?

- Since physical features is a function of species (most likely) and delta values depend upon the species, we should probably observe that physical features cluster around the similar delta values for a given species

```
## plot flipper length vs delta 15 values

ggplot(df, aes(x = delta_15_n_o_oo ,
               y = flipper_length_mm, color = common_name)) +
  geom_point() +
  facet_grid(common_name ~ .) + t
```

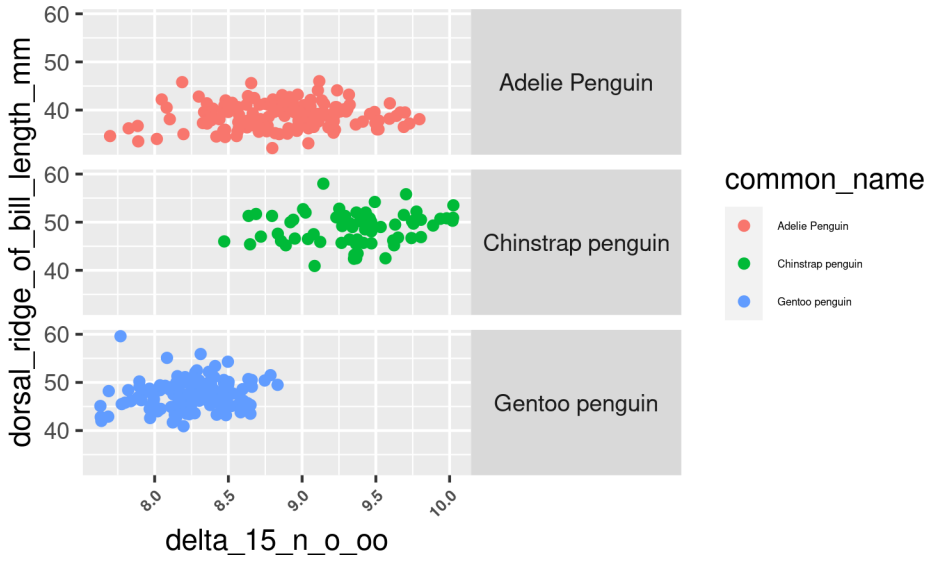



```
## plot flipper length vs delta 13 values
ggplot(df, aes(x = delta_13_c_o_oo ,
               y = flipper_length_mm, color = common_name)) +
  geom_point() +
facet_grid(common_name ~ .) + t
```

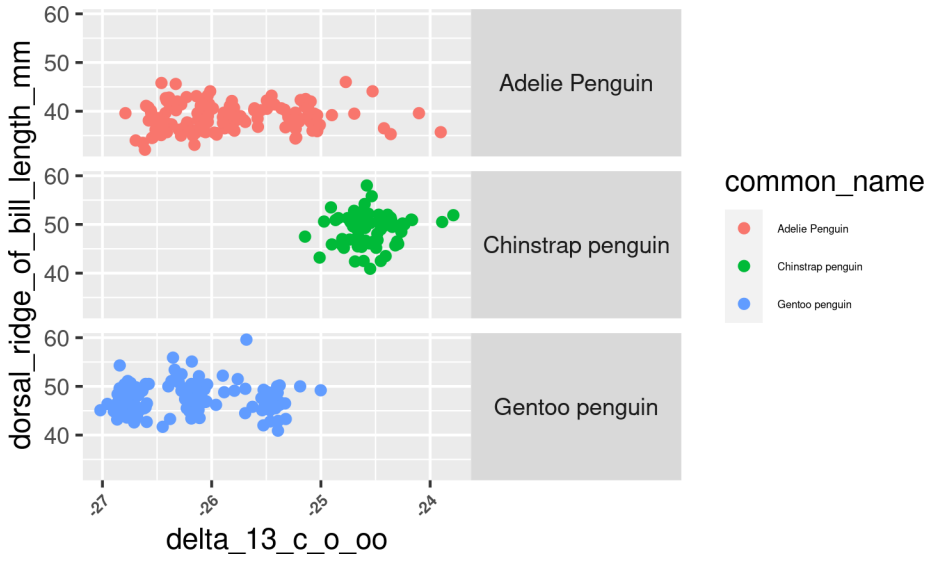



```
## plot dorsal_ridge_of_bill_length_mm vs delta 15 values
df %>% ggplot( aes(x = delta_15_n_o_oo,
               y =  dorsal_ridge_of_bill_length_mm, color = common_name)) +
```

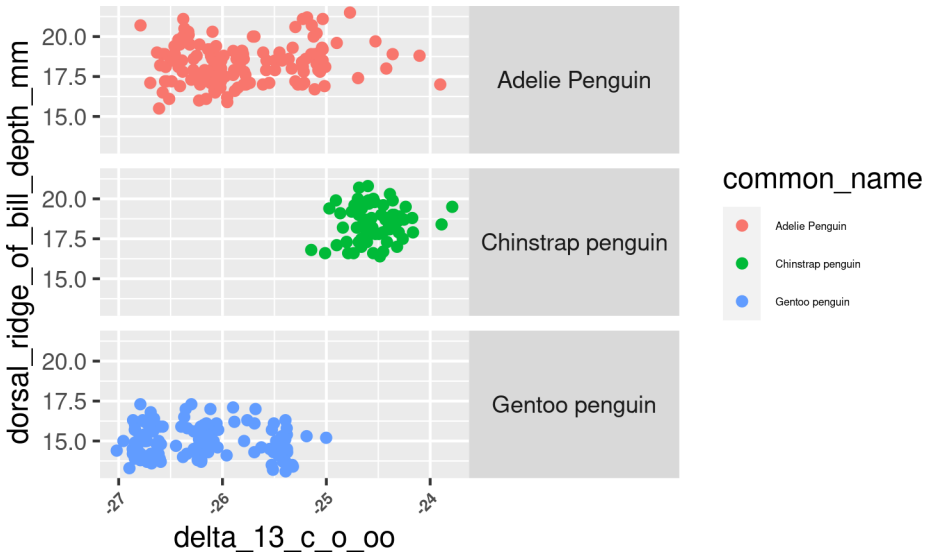
```
  geom_point() +
facet_grid(common_name ~ .) + t
```



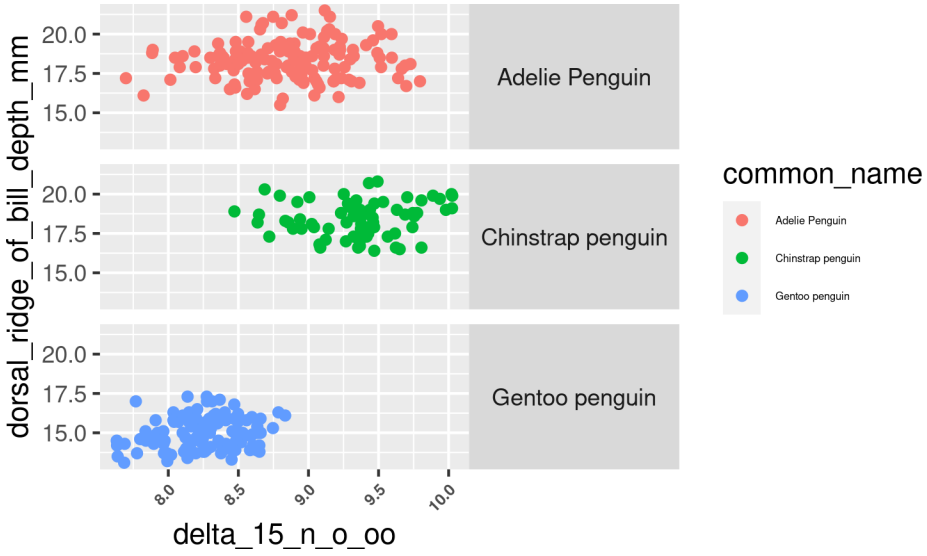

```
## plot dorsal_ridge_of_bill_length_mm vs delta 13 values
df %>% ggplot( aes(x = delta_13_c_o_oo,
                   y =  dorsal_ridge_of_bill_length_mm, color = common_name)) +
  geom_point() +
facet_grid(common_name ~ .) + t
```




```
## plot dorsal_ridge_of_bill_depth_mm length vs delta 13 values
df %>% ggplot( aes(x = delta_13_c_o_oo,
                   y =  dorsal_ridge_of_bill_depth_mm, color = common_name)) +
  geom_point() +
facet_grid(common_name ~ .) + t
```




```
## plot dorsal_ridge_of_bill_depth_mm length vs delta 15 values
df %>% ggplot( aes(x = delta_15_n_o_oo,
                   y =  dorsal_ridge_of_bill_depth_mm, color = common_name)) +
  geom_point() +
facet_grid(common_name ~ .) + t
```



5.4 Links

https://www.statsandr.com/blog/how-to-perform-a-one-sample-t-test-by-hand-and-in-r-test-on-one-mean/#assumptions

https://www.statsandr.com/blog/student-s-t-test-in-r-and-by-hand-how-to-compare-two-groups-under-different-scenarios/

http://www.r-project.org

http://rmarkdown.rstudio.com/