# Ocean Data Analysis with R Programming for Early Career Ocean Professionals (ECOPs) (Asia)

Univariate Statistical Tests

Mohamad Lukman Aidid bin Mohd Yusoff

2023-11-01

Lesson 2: Univariate statistical tests

1. We want to see if the temperature is significantly different depending on the country. In the case where all the assumptions are validated, which test will you perform?
2. Check the assumptions. What can you conclude?
3. Run the test that you think is most appropriate for this type of data. You can use the decision tree that was presented in the last lesson. What is th p-value? What can you conclude?

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)
library(ggplot2)
library(stats)
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
setwd('C:/Users/Administrator/Desktop/R/')
```

```r
obis <- read.csv("C:/Users/Administrator/Desktop/R/obis_red_list_filtered_1000.csv")
head(obis)
```

```
##          scientificName date_year          family minimumDepthInMeters
## 1 Balaenoptera physalus      2003 Balaenopteridae                    0
## 2 Balaenoptera physalus      2003 Balaenopteridae                    0
## 3 Balaenoptera physalus      2003 Balaenopteridae                    0
## 4 Balaenoptera physalus      2003 Balaenopteridae                    0
## 5 Balaenoptera physalus      2003 Balaenopteridae                    0
## 6 Balaenoptera physalus      2002 Balaenopteridae                    0
##   shoredistance   sst   sss individualCount    country status
## 1        182964 -1.47 34.03               2 Antarctica     VU
## 2        135623 -1.58 34.01               2 Antarctica     VU
## 3        138638 -1.58 34.01               9 Antarctica     VU
## 4         77966 -1.57 34.06               4 Antarctica     VU
## 5        141441 -1.59 34.02               3 Antarctica     VU
## 6        -14124 -1.43 33.71               3 Antarctica     VU
```

```r
unique(obis$country)
```

```
##  [1] "Antarctica"        "Australia"         "Spain"
##  [4] "United States"     "French Polynesia"  "Colombia"
##  [7] "Tonga"             "Papua New Guinea"  "Taiwan"
## [10] "The Netherlands"   "Bahamas"           "Cook Islands"
## [13] "Wallis and Futuna" "Fiji"              "Marshall Islands"
```

**Question 1:**

'Temperature' is a continuous data and 'country' is a categorical data (factor) with more than 2 groups. Thus, if all the assumptions are validated (namely normality, homoscedasticity, and independence), then the most appropriate test would be ANOVA.

**Question 2:**

Based on the the previous lesson assignment results, the data for 'temperature' is not normal and could not be transformed into a normal distribution via log transform, as both data did not pass the Shapiro-Wilk test .

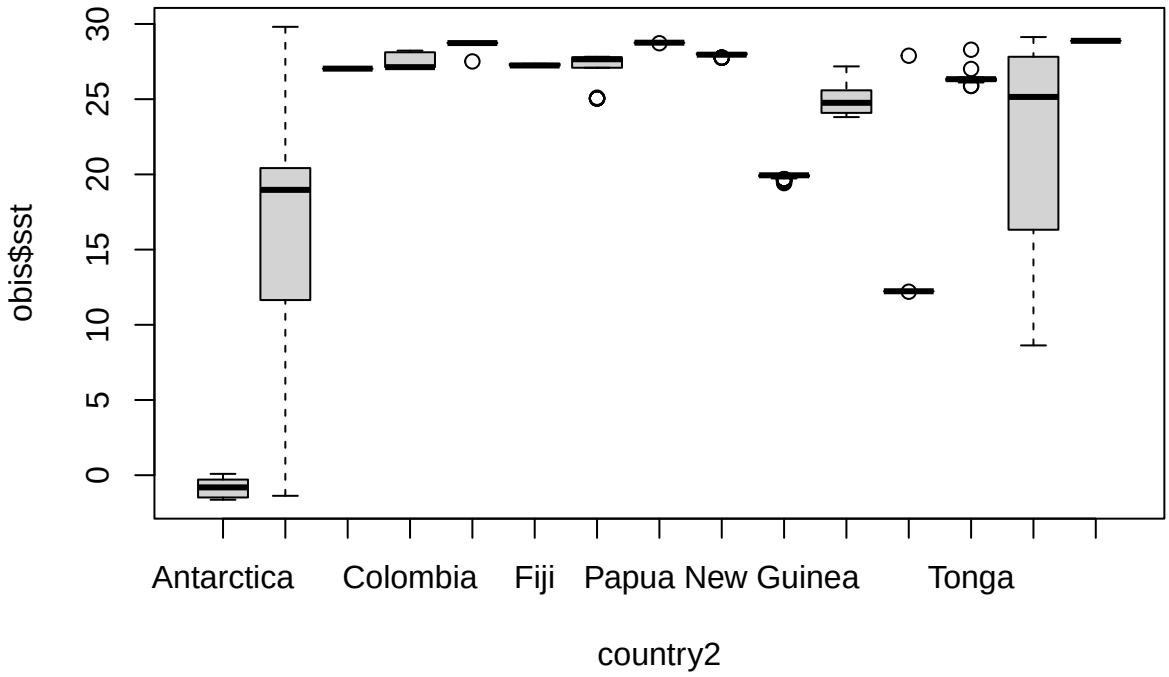
Comparing variances of 'temperature' and 'country' via the Levene test also indicated that the data set is not homoscedastic.

Therefore, ANOVA would not be suitable to test this data.

**Question 3:**

Based on assumptions put forth in question 2, the most appropriate test for this type of data would be the Kruskal-Wallis test.

```
country2 <- as.factor(obis$country)
plot(obis$sst~country2)
```




```
kruskal.test(sst~country, data=obis)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  sst by country
## Kruskal-Wallis chi-squared = 464.77, df = 14, p-value < 2.2e-16
```

As the p-value is less than the significance level 0.05, we can conclude that there are significant differences between the countries.