Guide to R

Online Resources

<https://www.statmethods.net> > tips and help on basic stats and graphing issues

<https://statmethods.wordpress.com> > a blog on common issues you may come across

<https://r4ds.had.co.nz> > if you need more complicated help

<https://ww2.coastal.edu/kingw/statistics/R-tutorials/> > hints for R codes

<https://stat.ethz.ch/~meier/teaching/anova/> > info on ANOVA

<https://www.rstudio.com/resources/cheatsheets/> > R studio cheat sheets

Tests

* 2 sample t-test (unpaired)
  + 2 numerical variables
  + Variables are randomly sampled
  + Distribution is normal
  + Variance is equal
  + Variables are independent
  + R CODE:
    - install.packages (“ggpubr”)
    - t.test(x, y, alternative = “two.sided”, var.equal = FALSE)
* Simple linear regression test
  + Used to model the relationship between two continuous variables (continuous variables are numbers containing decimals)
  + 2 variables, numerical or categorical, but one must be numerical
  + Distribution is normal
  + Variance is constant
  + Variables are independent
  + R CODE:
    - library(tidyverse)
    - library(ggpubr)
    - theme\_set(theme\_pubr())
    - ggplot(nameofdataset, aes(x = nameofxdata, y = nameofydata)) + geom\_point() + stat\_smooth() 🡪 this should give you a scatterplot with an estimated linear regression line
    - lm1 <- Lm(variable1~variable2, data = nameofdataset) 🡪 this will give you the coefficients and save it in your data as lm1
    - ggplot(nameofdataset, aes(x = nameofxdata, y = nameofydata)) + geom\_point() + stat\_smooth(lm1 = lm) 🡪 this will give you a scatterplot with the actual linear regression line
    - summary(lm1) 🡪 this will give you calculated residuals, coefficients, standard errors, R2, and f-statistic for the x variable
    - confint(lm1) 🡪 this will give you the confidence interval for this model
* One-way ANOVA (Analysis Of Variance)
  + Used to compare means in a situation of more than 2 dependent variables
  + For more than 2 variables (numerical or categorical), at least 3 populations
  + Populations must be independent
  + Variance in each group are the same
  + Variables must be independent from each other
  + Variables cannot overlap or influence other dependent variables
  + Each population is normally distributed
  + RCODE:
    - It is important that your data is attached to R because of the multiple variables. To do this, import your data in R then use the code attach(nameofdata)
    - Check levels of data: levels(nameofdata$group)
    - If levels are not in the correct order you can reorder them
      * Give your data a name: my\_data$group <- ordered(nameofdataset$group, levels = c(“level1”, “level2”, “level3”))
    - Download package: library(dplyr)
    - group\_by(nameofdataset, group)

summarise(

count = n(),

mean = mean(nameofvariable, na.rm = TRUE),

sd = sd(nameofvariable, na.rm = TRUE) 🡪 this will give you a table of stats

* + - install.packages(“ggpubr”)
    - ggboxplot(nameofdata, x = “x variable”, y = “y variable”, order = c(“level1”, “level2”, “level3”, ylab = “nameofxaxis”, xlab = “nameofxaxis”)
    - anova <- aov(xvariable ~ group, data = nameofdataset) 🡪 will compute the anova test
    - summary(anova) 🡪 will create a summary of the anova
    - Check for variance assumption: plot(anova, 1) 🡪 if red line is almost straight then the variances pass the assumption check

How to check for normal distribution

* qqnorm(nameofdataset$len, pch = 1, frame = FALSE) 🡪 this will create a normal QQ plot
* qqline(nameofdataset$len, lwd = 2) 🡪 this will add the reference line

How to check for equal variance (Levene’s test)

* library(car)
* leveneTest(xvariable ~ group, data = nameofdataset) 🡪 if p-value is not less than 0.05 then the variance passes the test

Boxplot

* ggboxplot(datasetname, x = “x variable data”, y = “y variable data”, ylab = “title of y axis”, xlab = “title of x axis”)