# Assignment 7 for Statistical Computing and Empirical Mathods

Dr. Henry WJ Reeve

Teaching block 1 2021

## Introduction

This document describes your sixth assignment for Statistical Computing and Empirical Methods (Unit EMATM0061) on the MSc in Data Science. Before starting the assignment it is recommended that you first watch video lectures 15, 16, 17.

Begin by creating an Rmarkdown document with html output. You are not expected to hand in this piece of work, but it is a good idea to get used to using Rmarkdown.

## 1 Student's t-confidence intervals

In this problem we will discuss a parametric approach to obtaining confidence intervals based upon Student's t-distribution. In the code below "adelie_flippers" is a vector containing the flipper lengths of a sample of Adelie penguins. The following code computes confidence intervals based on "adelie_flippers" for the population mean of the flipper lengths for Adelie penguins using the Student's t-distribution method.

```
alpha<-0.05
sample_size<-length(adelie_flippers)
sample_mean<-mean(adelie_flippers)
sample_sd<-sd(adelie_flippers)
t<-qt(1-alpha/2,df=sample_size-1)
confidence_interval_l<-sample_mean-t*sample_sd/sqrt(sample_size)
confidence_interval_u<-sample_mean+t*sample_sd/sqrt(sample_size)
confidence_interval<-c(confidence_interval_l,confidence_interval_u)
confidence_interval
```

What would happen to the width of my confidence interval if the sample mean were higher? What would happen to the width of my confidence interval if the sample standard deviation were higher? What would happen to the width of my confidence interval if the sample size were larger?

Use your data wrangling skills to extract a vector consisting of the weights of all the Red-Tailed hawks from the "Hawks" data set, with any missing values removed.

Now use the Student's t method to compute $99\%$-**level** confidence intervals for the population mean of the weights for the red tailed hawks. Note that opting for confidence intervals with a confidence level of $99\%$, rather than a confidence level of $95\%$, requires a modified value of $\alpha$.

What assumptions are made to derive confidence intervals based on Student's t-distribution? Check if these assumptions are justified using a kernel density plot with the **geom_density()** function and using a QQ-plot with the **stat_qq()** function.

## 2 One sample t-test

Begin by loading the "Palmer penguins" library. Next extract a vector called "bill_adelie" consisting of the bill lengths of the Adelie penguins belonging to the Adelie species.

Carry out a statistical hypothesis test to test the hypothesis that the population mean of the Adelie penguin's bill lengths is $40$mm. Use a significance level of $0.01$. You can use the **t.test()** function. What assumptions are required for this hypothesis test?

## 3 Implementing a one-sample t-test

Implement a function carries out a two sided one-sample t-test. Your sample should take in two arguments 1) a vector $x$ corresponding to a sample $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ and a 2) the value $\mu_0$ corresponding to a null hypothesis of $\mu = \mu_0$. The output of your function should be the corresponding $p$-value of the test.

You can test your implementation by confirming your function gives the same p-value as the **t.test()** function for the example in question 2 of the assignment.

## 4 The paired t-test

The Barley data set gives the yields of two types of barley - Glabron and Velvet across twelve different fields. The data is paired as yields are given for both types of barley across each of the twelve fields.

```
library(PairedData)
data("Barley")
```

Carry out a paired t-test to determine whether there is a difference in average yield between the two types of barley. Use a significance level of $0.01$. You can use the **t.test()** function.

Compute the effect size using Cohen's d statistic.

What assumptions are required for the one-sample t test? Are these assumptions justified in this case?

## 5 Investigating coverage for Student's t intervals

In this question we shall assume that we have access to a sample $X_1, \ldots, X_n \sim \mathcal{N}(\mu_0, \sigma_0^2)$ consisting of i.i.d. Gaussain data. We are interested in determining the value of the unknown population mean $\mu_0$ based upon the sample $X_1, \ldots, X_n$.

Suppose we wish to compute confidence intervals for $\mu_0$ with confidence level $(1-\alpha)\times 100\%$, for some $\alpha \in (0, 1)$. For example, we could have $\alpha = 0.05$, in which cases we wish to compute confidence intervals with confidence

level $95\%$. Let $\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean and $S := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$ be the sample standard deviation. In addition, let $t_{\alpha/2,n-1}$ be the $(1-\frac{\alpha}{2})$-quantile of the Student's t-distribution with $n-1$ degrees of freedom.

The Student's t confidence interval for $\mu_0$ is given by $(L(X_1,\ldots,X_n), U(X_1,\ldots,X_n))$ defined by

$$L_\alpha(X_1,\ldots,X_n) := \overline{X} - \frac{t_{\alpha/2,n-1}}{\sqrt{n}} \cdot S$$
$$U_\alpha(X_1,\ldots,X_n) := \overline{X} + \frac{t_{\alpha/2,n-1}}{\sqrt{n}} \cdot S.$$

The following code generates a function `student_t_confidence_interval`, which takes as input a sample $X_1,\ldots,X_n$ given as a vector along with a confidence level $\gamma = 1-\alpha$ and outputs a tuple containing $L_\alpha(X_1,\ldots,X_n)$ and $U_\alpha(X_1,\ldots,X_n)$:

```
student_t_confidence_interval<-function(sample,confidence_level){

  sample<-sample[!is.na(sample)] # remove any missing values
  n<-length(sample) # compute sample size
  mu_est<-mean(sample) # compute sample mean
  sig_est<-sd(sample) # compute sample sd
  alpha = 1-confidence_level # alpha from gamma
  t<-qt(1-alpha/2,df=n-1) # get student t quantile
  l=mu_est-(t/sqrt(n))*sig_est # lower
  u=mu_est+(t/sqrt(n))*sig_est # upper

  return(c(l,u))

  }
```

Check that you understand this function and implement it for yourself.

The key property of a confidence interval for $\mu_0$ at the confidence level of $(1-\alpha)\times 100\%$ is the following coverage property:

$$\mathbb{P}\{L_\alpha(X_1,\ldots,X_n) \leq \mu_0 \leq U_\alpha(X_1,\ldots,X_n)\} \geq 1-\alpha.$$

This is known as a coverage property since it tells us that the confidence interval *covers* $\mu_0$ with probability $1-\alpha$. The following simulation checks this property with $\mu_0 = 1$, $\sigma_0 = 3$ and a confidence level of $95\%$ i.e. $\gamma = 0.95$.

```
num_trials<-100000
sample_size<-30
mu_0<-1
sigma_0<-3
alpha<-0.05

set.seed(0) # set random seed for reproducibility

single_alpha_coverage_simulation_df<-data.frame(trial=seq(num_trials))%>%
  mutate(sample=map(.x=trial,.f=~rnorm(n=sample_size,mean=mu_0,sd=sigma_0)))%>%
  # generate random Gaussian samples
  mutate(ci_interval=map(.x=sample,.f=~student_t_confidence_interval(.x,1-alpha)))%>%
  # generate confidence intervals
  mutate(cover=map_lgl(.x=ci_interval,
```

```
                        .f=~((min(.x)<=mu_0)&(max(.x)>=mu_0))))%>%
  # check if interval covers mu_0
  mutate(ci_length=map_dbl(.x=ci_interval,
                        .f=~(max(.x)-min(.x))))
  # compute interval length

single_alpha_coverage_simulation_df%>%
  pull(cover)%>%
  mean() # estimate of coverage probability
```

```
## [1] 0.95003
```

Check that you understand the above code. Now modify the above code to conduct a simulation experiment to investigate how $\mathbb{P}\{L_\alpha(X_1,\ldots,X_n) \leq \mu_0 \leq U_\alpha(X_1,\ldots,X_n)\}$ varies as a function of the confidence level $\gamma = 1-\alpha$.

How does the average length $\mathbb{E}(|U_\alpha(X_1,\ldots,X_n) - U_\alpha(X_1,\ldots,X_n)|)$ vary as a function of the confidence level $\gamma = 1-\alpha$?

# 6 (Optional) Wilson's confidence interval for proportions

The following code uses Wilson's method to compute $99\%$-level confidence intervals for the pass rate of a driving test.

```
library(PropCIs)
```

```
driving_test_results<-c(1,0,1,0,0,0,0,0,0,1,0,0,0,1,0,1,0,1,0,1,0,0,1,0)
alpha<-0.01 # failure probability
num_successes<- sum(driving_test_results) # total passes
sample_size<-length(driving_test_results)
scoreci(x=num_successes, n=sample_size, conf.level=1-alpha)
# compute Wilson's  confidence intervals
```

Use Wilson's method to compute a $95\%$-level confidence interval for the proportion of red-tailed hawks who weigh more than a kilogram.

# 7 (Optional) The Binomial test

The "Airlines" data set contains arrival records for LaGuardia and O'Haire airport. We can load the "Airlines" test as follows:

```
library(Stat2Data)
data("Airlines")
```

Extract a subset of the data set corresponding to arrivals of flights with the Delta airline at the O'Hare airport.

Carry out a statistical hypothesis test to test the hypothesis that $87.5\%$ of the arrivals of flights with the Delta airline at the O'Hare airport are on time. Use a significance level of $0.05$. You can use the **binom.test()** function. What assumptions are required for this hypothesis test?

## 8 (Optional) Bootstrap confidence intervals

The following code computes a $95\%$-level confidence interval for the mean weight of the penguins.

```
library(boot) # load the library
set.seed(123) # set random seed

#first define a function which computes the mean of a column of interest
compute_mean<-function(df,indicies,col_name){
  sub_sample<-df%>%slice(indicies)%>%pull(all_of(col_name)) # extract subsample
return(mean(sub_sample,na.rm=1))}# return median

# use the boot function to generate the bootstrap statistics
results<-boot(data = penguins,statistic =compute_mean,col_name="body_mass_g",R = 1000)

# compute the 95%-level confidence interval for the mean
boot.ci(boot.out = results, type = "basic",conf=0.95)
```

Explain the importance of the random seed. What assumptions underpin this method?

Compute a $99\%$-level confidence interval for the median weight of the hawks using the Hawks data set.

What can we say about the relationship between the average Hawk weight and the average penguin weight?

## 9 (Optional) Investigating the failure probability for Wilson's method

This problem is a more challenging optional extra. Conduct a simulation study based on Bernoulli samples $X_1, \cdots, X_n \sim \mathcal{B}(q)$ with $n = 100$ and $q = 0.5$. Wilson's method generates a pair $[\hat{L}_{n,\alpha}(X_1, \cdots, X_n), \hat{U}_{n,\alpha}(X_1, \cdots, X_n)]$ so that for a given failure probability $\alpha$, we have

$$\mathbb{P}\left[\hat{L}_{n,\alpha}(X_1, \cdots, X_n) \leq q \leq \hat{U}_{n,\alpha}(X_1, \cdots, X_n)\right] \approx 1 - \alpha.$$

This approximation is based on the central limit theorem. Conduct a simulation study to investigate how the probability $\mathbb{P}[\hat{L}_{n,\alpha}(X_1, \cdots, X_n) \leq q \leq \hat{U}_{n,\alpha}(X_1, \cdots, X_n)]$ depends upon $\alpha$.

## 10 (Optional) Effect size for the one sample t-test

In this question we introduce a natural measure of effect size for the one sample t-test considered in question 2. Suppose we have a sample $X_1, \cdots, X_n$. We assume that $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are sampled independently and identically distributed from a Gaussian distribution. Our null hypothesis is $\mu = \mu_0$ and our alternative is $\mu \neq \mu_0$ for some $\mu_0 \in \mathbb{R}$. Suppose that our $p$-value is sufficiently small that we are justified in rejecting the null.

We conclude that $\mu \neq \mu_0$. The Cohen's d-statistic for the one sample t-test is computed as follows:

$$\hat{d} = \frac{\overline{X} - \mu_0}{S_X},$$

where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$.

Create a function called "effect_size_one_sample_t_test" which computes Cohen's d statistic for the one sample t-test. The function should have two arguments: The first input is a vector $x$ corresponding to the sample $X_1, \cdots, X_n$. The second "mu" corresponds to the quantity $\mu_0$ in the null hypothesis.

In question 2 we carried out a one sample t-test do test the hypothesis that the population mean of the Adelie penguin's bill lengths is $40$mm. We rejected the null at a significance level of $0.01$. Apply your function "effect_size_one_sample_t_test" to compute the effect size. Comment on the magnitude of the effect.

# 11 (Optional) Confidence intervals for the exponential distribution

Suppose that $X_1, \ldots, X_n \sim f_{\lambda_0}$ are independent and identically distributed random variables with exponential distribution and probability density function

$$f_{\lambda_0}(x) = \begin{cases} \lambda_0 e^{-\lambda_0 x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

Given $\alpha \in [0,1]$ let $z_{\alpha/2}$ be the $(1-\alpha/2)$-quantile of a standard Gaussian random variable $Z \sim \mathcal{N}(0,1)$. We can compute an (approximate) confidence interval with a confidence level of $(1-\alpha)\times 100\%$ for $\lambda_0$ as follows:

$$L_\alpha(X_1, \ldots, X_n) := \frac{1}{\overline{X}}\left(1 - \frac{z_{\alpha/2}}{\sqrt{n}}\right)$$
$$U_\alpha(X_1, \ldots, X_n) := \frac{1}{\overline{X}}\left(1 + \frac{z_{\alpha/2}}{\sqrt{n}}\right).$$

Create a function which takes as inputs a vector called `sample` consisting of i.i.d. exponentially distributed random variables $X_1, \ldots, X_n \sim f_{\lambda_0}$, and a confidence level `confidence_level` ($\gamma = 1-\alpha$) and outputs a confidence interval $(L_\alpha(X_1, \ldots, X_n), L_\alpha(X_1, \ldots, X_n))$ for the parameter $\lambda_0$ with confidence level $\gamma = 1-\alpha$.

Next conduct a simulation study to explore $\mathbb{P}\left(L_\alpha(X_1, \ldots, X_n) \leq \lambda_0 \leq U_\alpha(X_1, \ldots, X_n)\right)$ as a function of the confidence level $\gamma = 1-\alpha$.

Recall that population mean $E(X) = \frac{1}{\lambda_0}$ and population variance $\mathrm{Var}(X) = \frac{1}{\lambda_0^2}$. Use the central limit theorem to show that

$$\mathbb{P}\left(L_\alpha(X_1, \ldots, X_n) \leq \lambda_0 \leq U_\alpha(X_1, \ldots, X_n)\right) \approx 1 - \alpha,$$

when the sample size $n$ is very large.