Assignment 10 for Statistical Computing and Empirical Methods: Confidence intervals

Henry Reeve

Introduction

This document describes your eighth 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 lecture 10 entitled “Confidence intervals”.

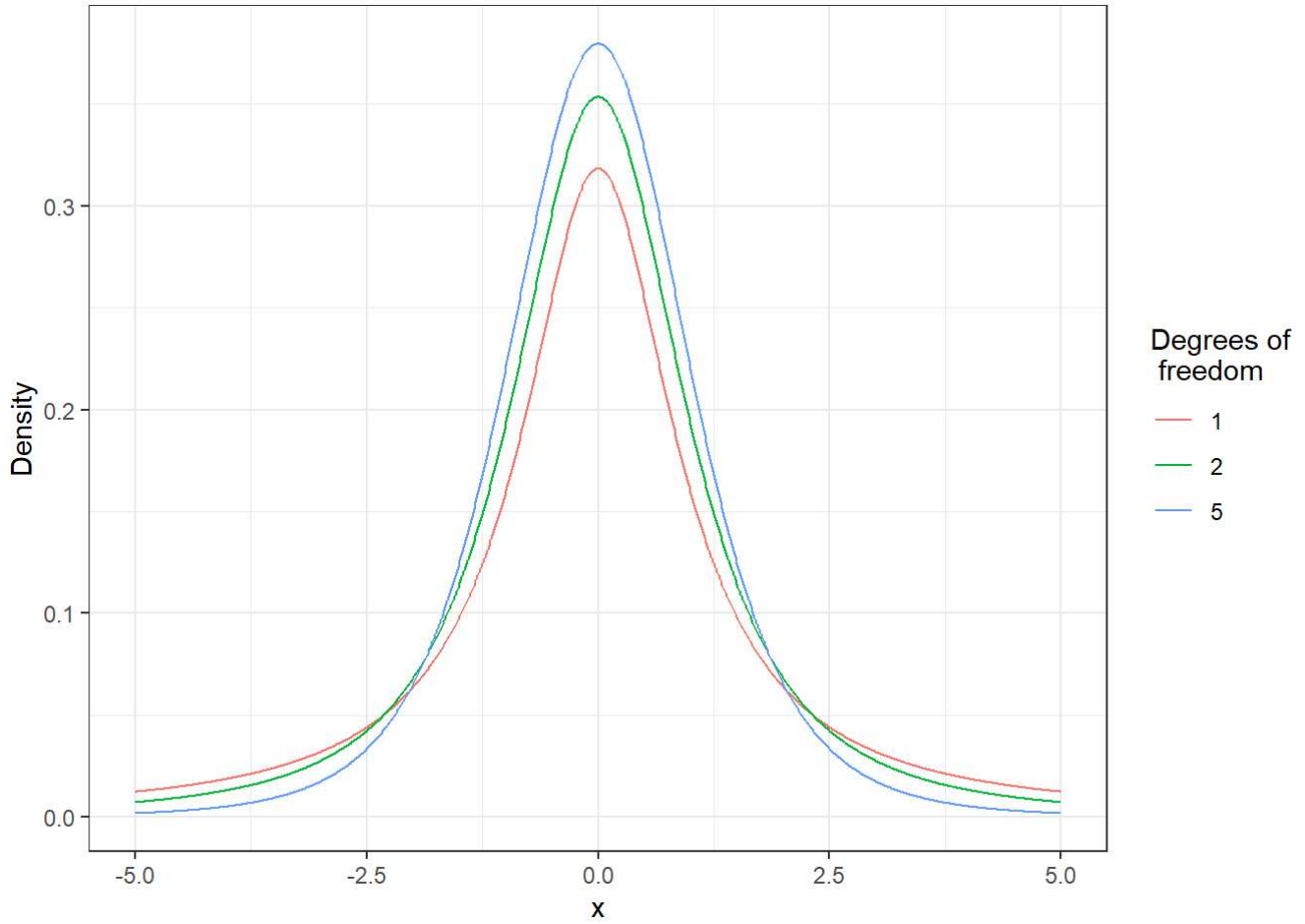
1 Student’s t-distribution

Student’s t-distribution plays a crucial role in deriving confidence intervals for the mean. Use the **dt()** function to generate density plots displaying Student’s t distributions with $k = 1,2$ and 5 degrees of freedom. Your plot should something like the following:

```
student_t_plot<-data.frame(x=numeric(),Density=numeric(),dof=character())

x=seq(-5,5,0.01)
for(dof in c(1,2,5)){
  student_t_plot<-student_t_plot%>%
    rbind(data.frame(
            x=x,Density=dt(x,df=dof),dof=as.character(dof)))
}


student_t_plot%>%
  ggplot(aes(x=x,y=Density,color=dof))+geom_line()+theme_bw()+
  labs(color="Degrees of\n freedom")
```



Explain the following output:

```
pt(qt(seq(0,1,0.1),df=3),df=3)
```

```
##  [1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
```

2 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.

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.

3 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?

4 Wilson's confidence interval for

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.

5 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 α, 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 α.