```
#Formula for the standard error of the mean
require(palmerpenguins)
head(penguins)
help(sd)
help("is.na")
sse_mean= function(x)
  {sse= sd(x, na.rm=TRUE)/(sqrt(length(x[! is.na (x)])))
  return(sse)
  }

sse_mean(penguins$bill_depth_mm)

#p-value is what allows us to determine if we can reject the null hypothesis

boxplot(flipper_length_mm ~ species, data=penguins)

#2 species data
dat_pen=subset(penguins, species !="Gentoo")
boxplot(flipper_length_mm ~ species, data=dat_pen)

dat_pen= droplevels(subset(penguins, species !="Gentoo"))
{
  par(mfrow=c(1,2))
  boxplot(flipper_length_mm ~ species, data=penguins)
  boxplot(flipper_length_mm ~ species, data=dat_pen)
}

#Resampling with replacement
  #for reproducibility
set.seed(123)

flipper_shuffled= sample(penguins$flipper_length_mm, replace=TRUE)
par(mfrow=c(1,2))
boxplot(flipper_length_mm ~ species, data=penguins)
boxplot(flipper_shuffled ~ penguins$species, xlab="species")
#the shuffled data breaks the connection between species and flipper length
#"The flipper lengths for the penguin species are drawn from the same
population
  #of all flipper lengths" aka there's no difference in length between species
#this is Monte Carlo resampling: breaks up associations and allows you to
  #simulate what would happen in the null hypothesis were true
  #great way to simulate a null distribution

#Bootstrap Resampling does not destroy associations in data, it samples entire
  #rows of your data. It shuffles rows with replacement but not columns

#Classical t-test
t.test(dat_pen$flipper_length_mm ~ dat_pen$species)
  #this t-test suggest that there is good evidence of different flipper
lengths
    #between the 2 species

#Two-sample resampling
set.seed(1)
```

```
flipper_shuffled= sample(dat_pen$flipper_length_mm)
boxplot(flipper_shuffled ~ dat_pen$species)

#Classical test on resampled data
t_test_1= t.test(flipper_shuffled ~ dat_pen$species)
t_test_1
  #t-test output does not support rejecting a null hypothesis that the 2
flipper
    #lengths are different between the 2 species

#Difference of means
t_test=t.test(dat_pen$flipper_length_mm ~ dat_pen$species)
t_test
t_test$estimate
diff_observed = round(diff(t_test$estimate), digits=3)
print(diff_observed, digits=3)

#Using aggregate()
  #allows you to calculate the difference in means
agg_means=aggregate(flipper_length_mm ~ species,
                    data=dat_pen, FUN=mean,
                    na.rm=TRUE)
diff_observed=diff(agg_means[, 2])
agg_means
diff_observed

#Sample sizes
  #tells the number of individuals of each species in the data
table(dat_pen$species)
  #resampling with replacement is the same thing as randomly sampling 68
flipper
    #lengths in one group and 152 in another
n_1=68
n_2=152

dat_1= sample(dat_pen$flipper_length_mm, n_1, replace = TRUE)
dat_2= sample(dat_pen$flipper_length_mm, n_2, replace = TRUE)

diff_simulated= mean(dat_1, na.rm = TRUE) - mean(dat_2, na.rm = TRUE)
print(c(observed = diff_observed, simulated= diff_simulated))

#Simulation function
two_group_resample= function(x, n_1, n_2)
{
  x=dat_pen$flipper_length_mm
  n_1=68
  n_2=152

  dat_1=sample(x, n_1, replace=TRUE)
  dat_2=sample(x, n_2, replace=TRUE)

  diff_simulated=mean(dat_1, na.rm=TRUE) - mean(dat_2, na.rm=TRUE)
  return(diff_simulated)
  }
```

```
set.seed(54321)
two_group_resample(dat_pen$flipper_length_mm, 68, 152)
#My two_group_resample isn't giving me an output, had to add return() to get
it
  #to spit out the value

#Resampling Experiment
n=2000
mean_differences=c()
for(i in 1:n)
{
  mean_differences=c(
    mean_differences,
    two_group_resample(dat_pen$flipper_length_mm, n_1, n_2))
}
hist(mean_differences)

sum(abs(mean_differences) >= diff_observed)

#Retrieving named elements
  #use the str() function to see what an object contains, use the $ to
retrieve
    #the itmes of interest
t_test=t.test(flipper_shuffled ~ dat_pen$species)
str(t_test)

t_test$estimate


##LAB QUESTIONS
#Q1
rm(list=ls())

sse_mean= function(x)
{sse= sd(x, na.rm=TRUE)/(sqrt(length(x[! is.na (x)])))
return(sse)
}

sse_mean(penguins$body_mass_g)
sse_mean(mtcars$mpg)

#Q2
two_group_resample= function(x, n_1, n_2)
{ dat_1=sample(x, n_1, replace=TRUE)
  dat_2=sample(x, n_2, replace=TRUE)

  difference_in_means=mean(dat_1, na.rm=TRUE) - mean(dat_2, na.rm=TRUE)
  return(difference_in_means)
}

#Q4
n = 2000
mean_differences = c()
```

```
for (i in 1:n)
{
  mean_differences = c(
    mean_differences,
    two_group_resample(dat_pen$flipper_length_mm, 68, 152)
  )
}
hist(mean_differences, main = "Adelie and Chinstrap Flipper Length Mean
Differences",
     xlab ="Mean Differences")

#Q5
sum(abs(mean_differences) >5.8)

#Q7
boxplot(bill_length_mm ~ species, data=dat_pen,
        main="Bill Length by Species", xlab="Species", ylab="Bill Length mm")

#Q8
agg_means = aggregate(
 bill_length_mm ~ species,
  data = dat_pen,
  FUN = mean,
  na.rm = TRUE)
diff_crit = diff(agg_means[, 2])

agg_means
diff_crit

#Q9
t.test(dat_pen$bill_length_mm ~ dat_pen$species)


#Q10

n=10000
mean_differences=c()
for(i in 1:n)
{mean_differences=c(
  mean_differences,
      two_group_resample(dat_pen$bill_length_mm, 68, 152)
)}

sum(abs(mean_differences) > diff_crit)

#Q11
hist(mean_differences, main="Mean Differences in Bill Length\n Adelie and
Chinstrap",
     xlab="Mean Differences")
```