Unit 6 Homework: Tests and and Confidence Intervals

w203: Statistics for Data Science

Low-Oxygen Statistics

The file `expeditions.csv` contains data about 10,000 climbing expeditions in the Himalayan Mountains of Nepal. The data was compiled by the Himalayan Database and published in csv format on Tidy Tuesday.

First, navigate to https://github.com/rfordatascience/tidytuesday/tree/master/data/2020/2020-09-22 to read some basic information about the data and examine the codebook.

The variable `highpoint_metres` represents the highest elevation reached by each expedition. Your task is to test whether the mean highest elevation is above 7400 meters.

a. Using the documentation about the data, your background knowledge, and the data itself, assess whether the assumptions underlying a valid t-test are met. If plots are useful to make this argument, include them; if numeric statements are useful to make this argument, use them.

Answer

One sample test of means can use t-test to compare the mean of a sample to a pre-specified value and tests for a deviation from that value.

In general, one samples independent t-test assume the following characteristics about the data:

(1) Independence of the observations. It is YES in our case;
(2) No significant outliers in the two groups. By plotting the "Histogram for highpoint_metres" in next page, there were no extreme outliers found.
(3) Normality, i.e. the data for each group should be approximately normally distributed. The highpoint_metres in our case is found to be normal distribution at around the 7400 area, referring to the graphs in the following pages.

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

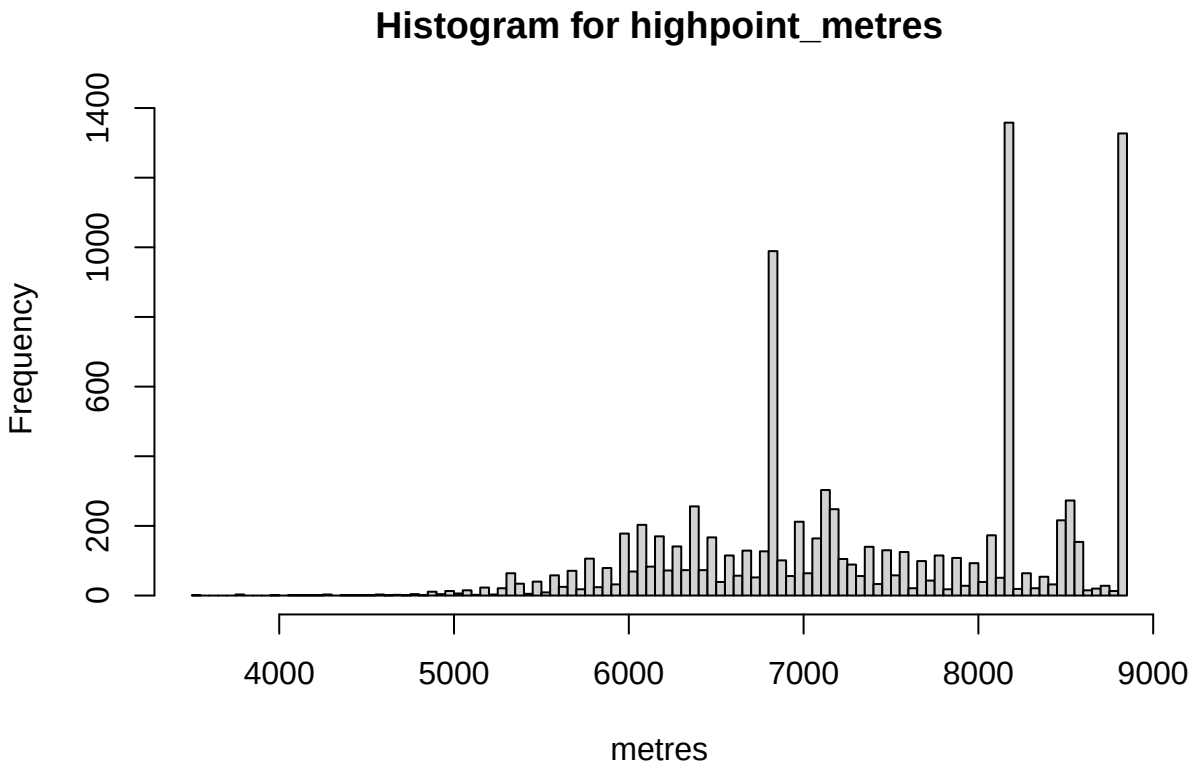
```
library(ggpubr)
library(rstatix)
```

```
##
## Attaching package: 'rstatix'

## The following object is masked from 'package:stats':
##
##     filter
```

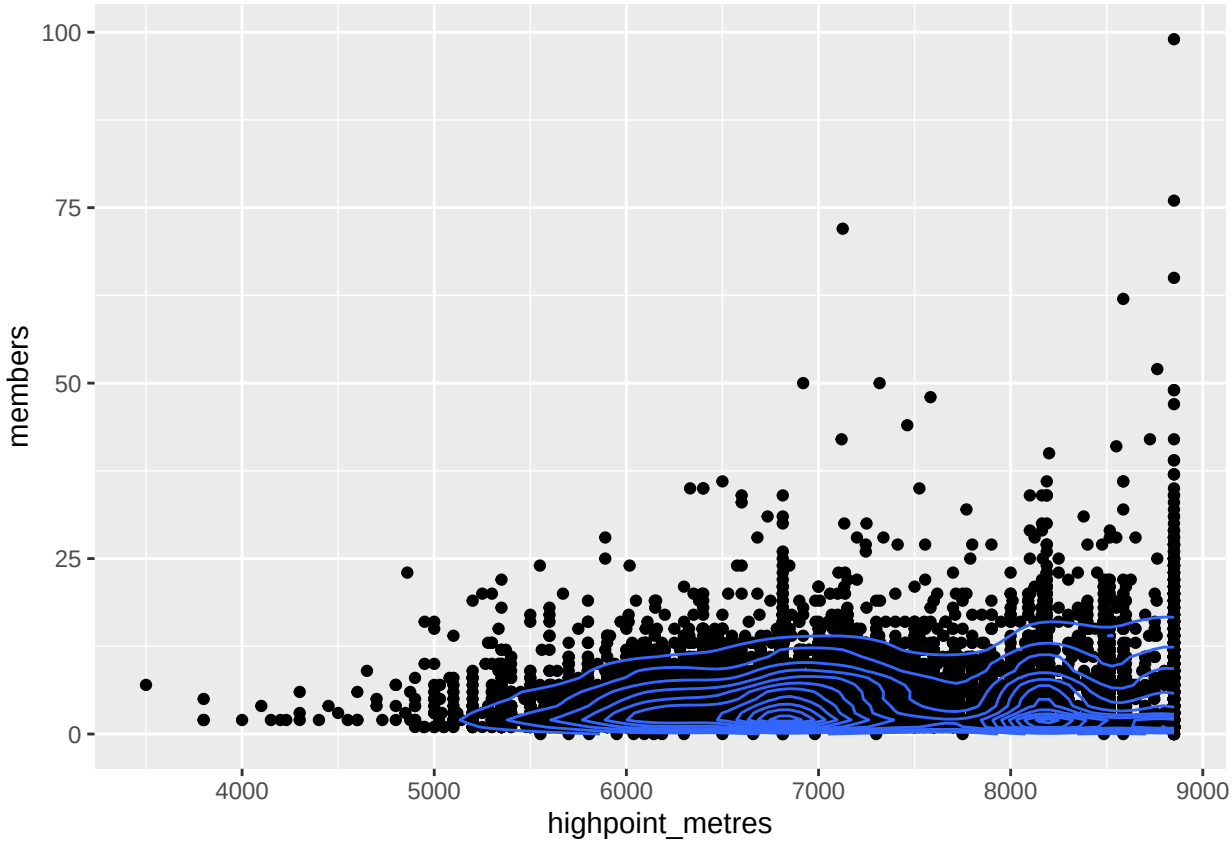
```
library(dplyr)
library(nortest)
library(splitstackshape)

# Load the data
e_read<-read.csv('expeditions.csv', header=TRUE)
e<-e_read%>%drop_na(highpoint_metres)

#e<-expandRows(e,"members")
metres<-e$highpoint_metres

summary(metres)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3500    6700    7300    7409    8188    8850
```

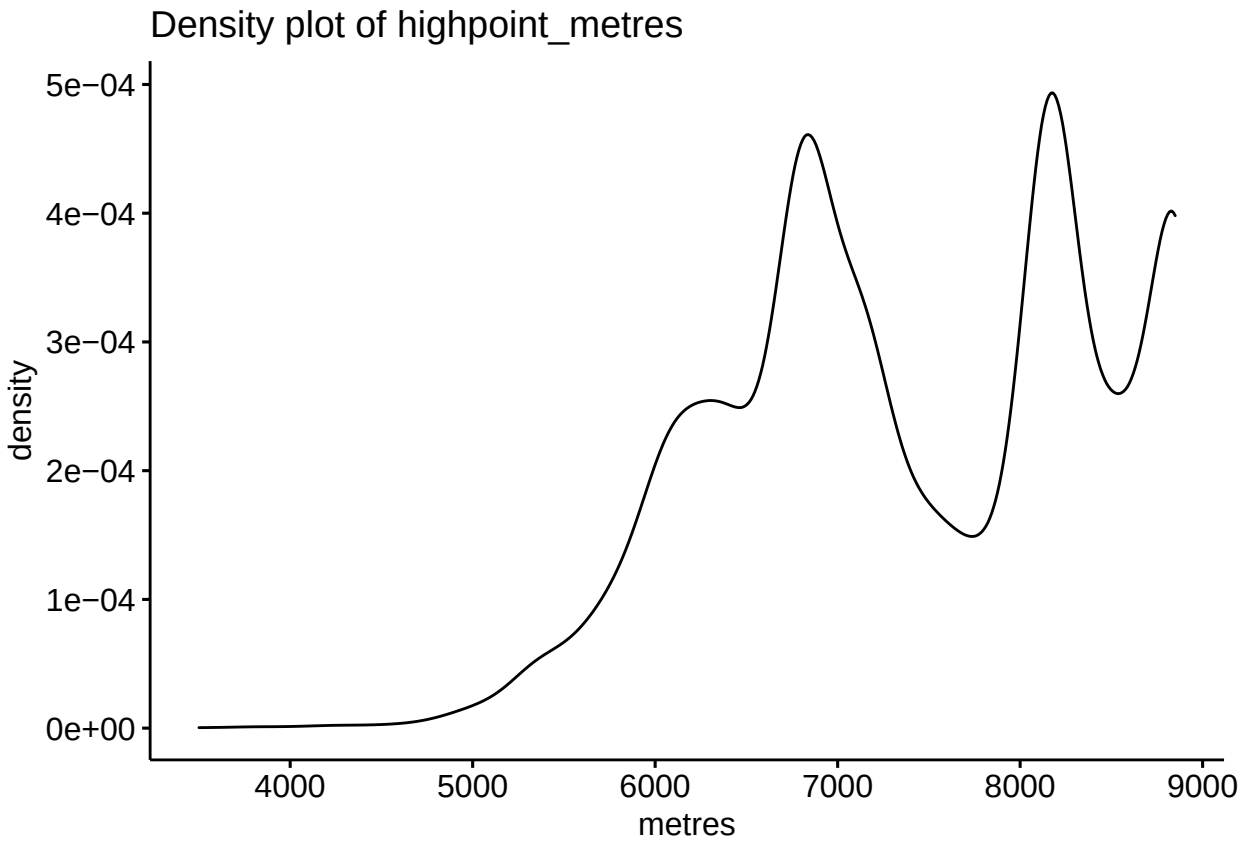
```
hist(metres,
  xlab = "metres",
  main = "Histogram for highpoint_metres",
  breaks = sqrt(nrow(e))
) # set number of bins
```



```
#Plot highpoint_metres vs members
ggplot(e, aes(x = highpoint_metres, y = members))+
  geom_point()+
  stat_density2d()
```



```
#Density plot
ggdensity(metres,
          main = "Density plot of highpoint_metres",
          xlab = "metres")
```



b. Provide an argument for why you should conduct a two-tailed test in this case, even though your personal interest is primarily in whether the mean is higher than 7400.

Answer

Here a two-tailed test is used because whether the mean is greater than or less than the target value, i.e. 7400, should be figured out. Two-tailed tests can test for effects in both directions. When performing a two-tailed test, the significance level percentage between both tails of the distribution is split.

One-tailed test is not used in this case because it is only justified if we have a specific prediction about the direction of the t-test, Or if we completely uninterested in the possibility that the opposite outcome could be true.

c. Compute the t-statistic by plugging in the values from the data manually into the formula. A great solution would write a function (perhaps called `t_statistic`)} **that takes arguments and returns a value. However, writing a function isn't necessary for a full solution. Feel free to use functions 'mean()', 'sd()', and 'sqrt()'.**

Answer

```
t_statistic <- function(highpoint_metres, mean_highest_elevation) {
   t <- (mean(highpoint_metres)-mean_highest_elevation)/
     (sd(highpoint_metres)/sqrt(length(highpoint_metres)))
   cat("t = ", t)
}

d<- e$highpoint_metre
h<-7400

t<-t_statistic(d, h)
```

```
## t =  0.8790473
```

d. Using 'qt()', compute the t-critical value for a two-tailed test. Answer

```
#df is degree of freedom

df<-length(metres)-1

df
```

```
## [1] 9949
```

```
#compute the t-critical value for a two-tailed test
t_critical_two_tailed<-qt(p=.05/2, df, lower.tail=FALSE)
cat("t_critical_two_tailed=", t_critical_two_tailed)
```

```
## t_critical_two_tailed= 1.960202
```

When perform a two-tailed test, there will be two critical values. In this case, the T critical values are 1.960202 and -1.960202. Thus, if the test statistic is less than -1.960202 or greater than 1.960202, the results of the test are statistically significant.

e. Compute the p-value for your two-tailed test. You may use the 'pt()' function.

Answer

```
t<-0.879043

p_value<-2*pt(-abs(t),df=length(metres)-1)

p_value
```

```
## [1] 0.3793992
```

f. Explain what your rejection decision should be in two ways.

Answer

If the P-value is less than (or equal to) 0.025, then the null hypothesis is rejected in favor of the alternative hypothesis. And, if the P-value is greater than 0.025, then the null hypothesis is not rejected.

In this case, the P_value is 0.3793992, no rejection.

g. Confirm that your work is correct, by running the 't.test' command.

Answer

Yes, it shows both of the results match each other, confirmed that the work is correct by running the `t.test` command.

```
#require(moonBook)
require(webr)
```

```
## Loading required package: webr
```

```
t.test(e$highpoint_metres, mu=7400, alternative = "two.sided")
```

```
##
##  One Sample t-test
##
## data:  e$highpoint_metres
## t = 0.87905, df = 9949, p-value = 0.3794
```

```
## alternative hypothesis: true mean is not equal to 7400
## 95 percent confidence interval:
##  7389.024 7428.823
## sample estimates:
## mean of x
##  7408.924
```

h. Evaluate the practical significance of your result.

Answer

The result is not practically significant, because from the data samples, we found most of people challenge some remarkable heights, the highest densities of heights are around 6800 metres, 8200 metres and 8800 metres.