Dhakal_Module2

Kokil Dhakal

2023-11-13

Question1

1.Summarize the data by whether children participated in the meal preparation or not.Use an appropriately labelled table to show the results.Also include a graphical presentation that shows the distribution of calories for participants vs.non-participants.Describe the shape of each distribution and comment on the similarity (or lack thereof) between the distributions in each group.

Answer:

loading the data

```
participants <- read_excel(path = "/Users/kokildhakal/Desktop/STUDY/BU/9.CS555/Module2/calorie_intake.xl
                           sheet ="participants")
non_participants <- read_excel(path = "/Users/kokildhakal/Desktop/STUDY/BU/9.CS555/Module2/calorie_intak
                           sheet ="non_participants")
```

making dataframe for each dataset type.

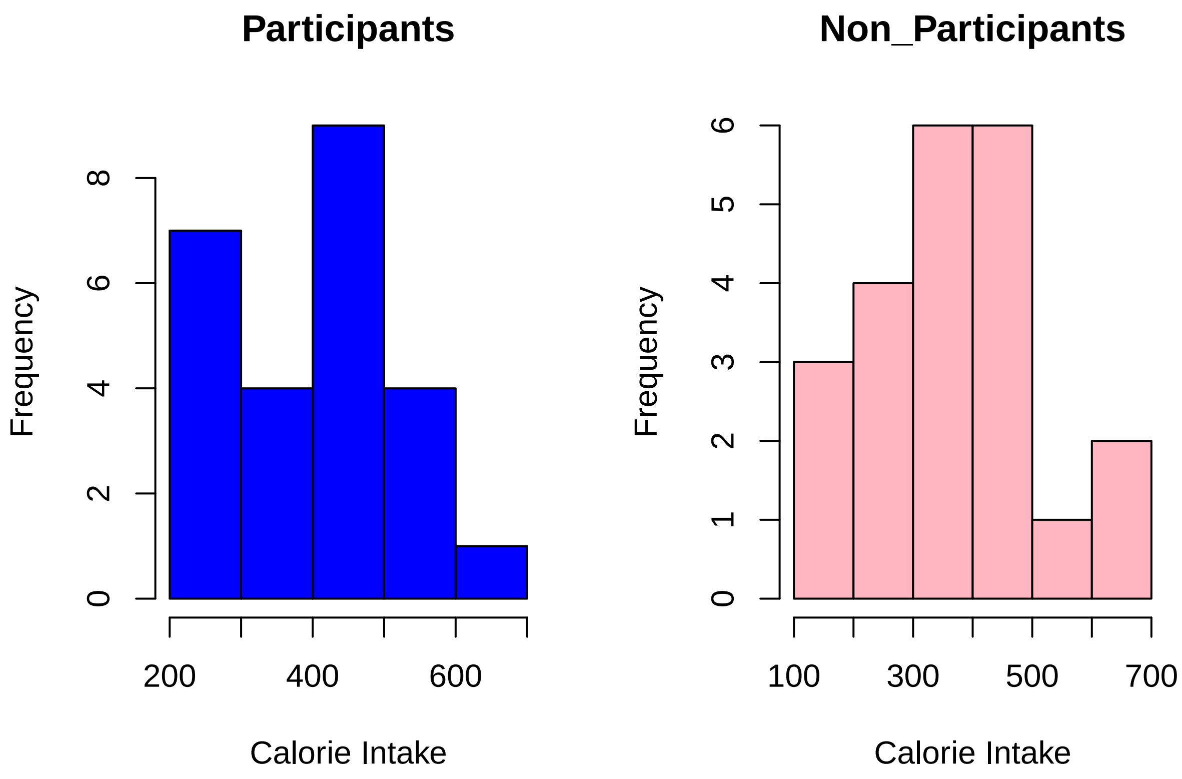
```
fivenum.participants <- fivenum(participants$`Calorie Intake for participants`)
participants.df <- data.frame(
  min=fivenum.participants[1],
  max=fivenum.participants[5],
  Q1=fivenum.participants[2],
  Q3=fivenum.participants[4],
  median=fivenum.participants[3],
  mean=mean(participants$`Calorie Intake for participants`),
  sd=sd(participants$`Calorie Intake for participants`),
  count=length(participants$`Calorie Intake for participants`),
  row.names = "participants"
)

fivenum.non_participants <- fivenum(non_participants$`Calorie intake for non-participants`)
non.participants.df <- data.frame(
  min=fivenum.non_participants[1],
  max=fivenum.non_participants[5],
  Q1=fivenum.non_participants[2],
  Q3=fivenum.non_participants[4],
  median=fivenum.non_participants[3],
  mean=mean(non_participants$`Calorie intake for non-participants`),
  sd=sd(non_participants$`Calorie intake for non-participants`),
  count=length(non_participants$`Calorie intake for non-participants`),
  row.names = "non_participants"
)

#combining each row from both data frame and summarize in a table.
combined.df <- rbind(participants.df,non.participants.df)
summarized.table <- kable(combined.df,format = "simple",align = "c",
                          caption = "Table of summary of two dataset",digits = 2)
summarized.table
```

Table 1: Table of summary of two dataset

	min	max	Q1	Q3	median	mean	sd	count
participants	210.99	635.21	298.38	456.30	424.94	410.08	121.51	25
non_participants	139.69	688.77	295.28	448.55	374.74	374.07	133.14	22

Now showing the distribution using the hist plotting

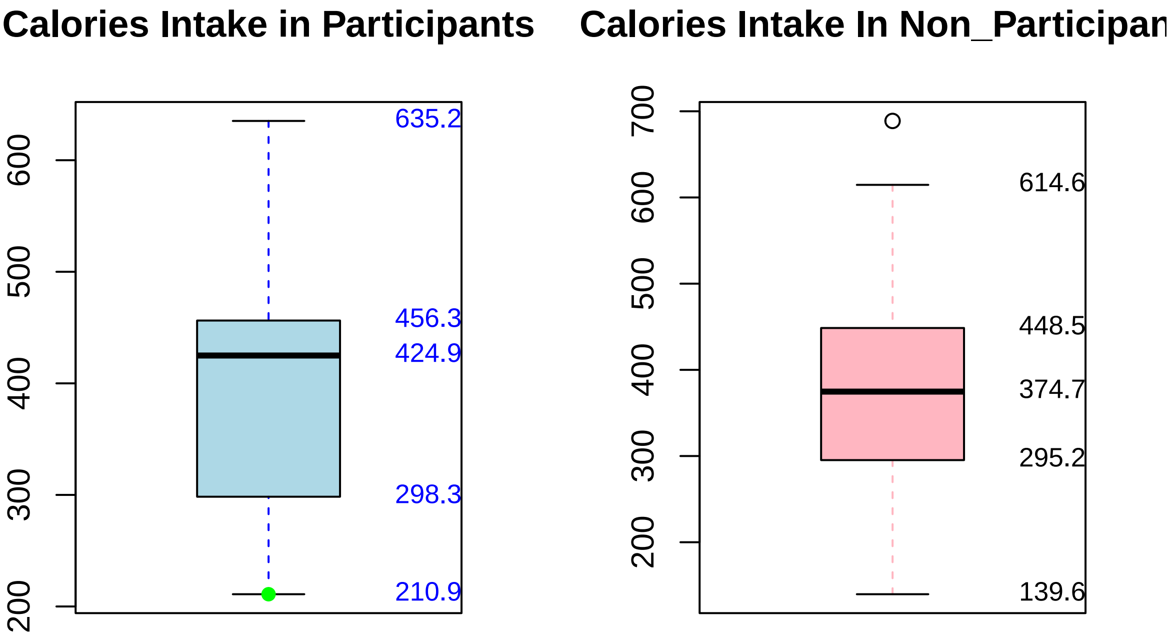
```
par(mfrow=c(1,2))
hist(participants$`Calorie Intake for participants`,main = "Participants",
     xlab = "Calorie Intake",col = "blue")
hist(non_participants$`Calorie intake for non-participants`,
     main = "Non_Participants",
     xlab = "Calorie Intake",col = "lightpink")
```



now showing the distribution using box plot

```
library(ggplot2)
stat.participants <- boxplot.stats(participants$`Calorie Intake for participants`)
stat.non.participants <- boxplot.stats(non_participants$`Calorie intake for non-participants`)
```

```
par(mfrow=c(1,2))
boxplot(participants$`Calorie Intake for participants`,main = "Calories Intake in Participants",outline
points(stat.participants$stats, col="green",pch=16)
text(x=1.3,y=sort(stat.participants$stats),labels = as.character(stat.participants$stats),pos = 4,col =

boxplot(non_participants$`Calorie intake for non-participants`,main = "Calories Intake In Non_Participar
text(x=1.3,y=sort(stat.non.participants$stats),labels = as.character(stat.non.participants$stats),pos =
```



For this question,I copy and paste each of the dataset for participants and non-participants in two different sheet of the excel and make two different dataframe from two different sheet. and then making summary of each dataset with values like min, max,Q1,Q3,mean, median,sd and count. After that making a table using kable() method and each row of the table represent different dataset and columns represet different values of the dataset.At last plotting each dataset using histogram and boxplot side by side for each dataset for comparision.

After reviewing summary table and plots,distribution of calorie intake for both participants and non-participants are not perfectly symmetrical.non_participants has one outlier while participants dataset has no outlier.Mean and median of the non-participants are almost same which means this dataset is almost symmetrical and it can be seen from the boxplot as well.Looking into box plot, it can be seen that calories intake participants dataset is slightly right skewed.The dataset for participants are more spread as we can check it by calculationg inter qrtile range(IQR) (ie.157.92 vs 153.27).

Question 2

2.Does the mean calorie consumption for those who participated in the meal preparation differ from 425? Formally test at the alpha = 0.05 level using the 5 steps outlined in the module.

Answer:solving this question using five steps as follow.

step1: Setting up the hypotheses and select the alpha level

H0:µ=425 (mean calories consumption for those who participated in the meal preparation is not differ from 425)

H1:µ $\neq$ 425 (mean calories consumption for those who participated in the meal preparation is not 425).

alpha=0.05

step2: Selecting the appropriate test statistic

Since sample size is less than 30.Also,Since we are comparing a sample mean to a known value and we don't have information about the population standard deviation,we can use a one-sample t-test.

t=(x$^{-}-\mu$)/(s/sqrt(n))

where x^{-}=sample mean,μ=population mean under null hypothesis,s=sample sd,n=sample size.

step3: State the decision rule

Determining the appropriate critical value from the t-distribution.Since we use two-sided t-distribution and we have total alpha= 0.05 which means each side will have probability of alpha/2=0.05/2=0.025. Using software(qt(0.025,24,lower.tail = FALSE)),the appropriate critical value is 2.064

Decision Rule: Reject H0 if |t| $\geq$ 2.064 (or p-value $\geq$ 0.05)

Otherwise, do not reject H0

Step4: Compute the test statistic and the associated p-value

```
#z=(sample.mean-population.mean at null hypothesis)/sample.sd/sqrt of n
t.2=(410.08-425)/(121.51/sqrt(25))
t.2
```

```
## [1] -0.6139412
```

```
p=pt(t.2,24)
p*2 # for both side
```

```
## [1] 0.5450302
```

step5: Conclusion

Since absolute value of t is smallar than critical value at 0.05 significance level(2.064),there is not enough evidence to reject null hypothesis.This means mean calories consumption for those who participated in the meal preparation is 425(p=0.5450302).

Question 3

3.Calculate a 90% confidence interval for the mean calorie intake for participants in the meal preparation. Interpret the confidence interval.

Answer

To calculate a confidence interval with a confidence level of the population mean, we use the following formula:

```
## x¯± t*s/√n <- This is the formula for confidence interval
## from t table value of t at 90% confidence level with 25-1=24 of df is 1.711
t.90 <- qt(0.05,24,lower.tail = FALSE) #using software for two sided t-curve.
t.90
```

x¯± t.90*s/√n. where x¯=sample mean, t.90= critical value at 90% confidence level. s= sample standard deviation, and n= sample size i.e 25.

```
## [1] 1.710882
```

```
lower_limit <- 410.08-t.90*(121.51/sqrt(25)) # mean and sd are obtained from above table.
upper_limit <- 410.08+t.90*(121.51/sqrt(25))
cat("from this calculation we are 90% cofident that true mean of calorie intake in participants is betwe
```

```
## from this calculation we are 90% cofident that true mean of calorie intake in participants is betwee
```

For question number 3,since samle size is less than 30, we will be using t-distribution curve to determine the confidence interval.sample mean and sample sd of participants is obtained from above summary table.Confidence interval gives minimum and maximum values between which a population mean falls and we can confirmed it with given percentage of confidence.In our case here, we can be 90% sure that population mean of calories intake in participants is within 368.5021 and 451.6579.

Question 4

4.Formally test whether or not participants consumed more calories than non-participants at the alpha = 0.05 level using the 5 steps outlined in the module.

Answwer:

Testing the above question using 5 steps as outline in module 2.

here, lets assume 1 is the mean calories consumed by participants and 2 is the mean carolies consumed by non-participants.

step1: Selecting the hypothesis and select the alpha level

H0:$\mu 1 <= \mu 2$**(The mean calories consumed by non-participants is less than or equall to mean calories consumed by participants)**

H1:$\mu 1 > \mu 2$**(mean calorie consumed by participants are higher than mean calorie consumed by non-participants).**

α**=0.05**

step2: Selecting the appropriate test-statistic

t= (x1¯−x2¯)/√(s1^2 /n1 + s2^2 /n2) where x1¯,s1,n1 are mean,SD,sample size of calories intake by participants and x2¯,s2,n2 are mean,SD,sample size of calories intake by non-participants respectively.

step3: State the decision rule

Determining the appropriate critical value from the standard t-distribution table associated with a right hand tail probability of 0.05 or finding critical values using dt() with minimum degree of freedom min(n1-1,n2-1)=21.from table with 21 degree of freedom and with 0.05 gives critical value of 1.721.

Decision Rule: Reject H0 if |t| $\geq$ **1.721**

Otherwise, do not reject H0

step4: Computing the test statistic and the associated p-value

```
## t= (x1¯-x2¯)/√(s1^2/n1+s2^2/n2)
t.4 <- (410.08-374.74)/sqrt((121.51^2/25)+(133.14^2/22))
t.4
```

calculating t value from the equation in setp 2.

```
## [1] 0.9457429
```

```
#calculating p-value
pt(t.4,21,lower.tail = FALSE) # df=min(n1-1,n2-1)=21
```

```
## [1] 0.1775177
```

step5: Conclusion

From above calculation, t value is 0.9457.According to our decision rule, t value need to be greater than 1.721 to reject null hypothesis.Since t value is less than 1.721, we do not have sufficient evidence at alpha=0.05 to reject null hypothesis(p=0.1775177). which gives evidence that mean calories intake by participants is not higher than mean calories intake by non-participants.

Question 5

Are the assumptions of the test used in (4) met? How do you know?

answer

In the two sample test,we assume that

1.observations are independent from each other.

2.We also assume that the distribution of the parameter in the population of interest is normally distributed or approximately normally distributed.

3.we assume that the population mean is unknown and that we are interested in making conclusions about this parameter using data from our sample.

Here, first and third assumptions are believe to be met as we consider the observations are independent and population mean is unknown. In case of second assumption,as we can see from above histogram,it is approximately normally distributed.Also,for sample size less than 30 we adjust the normal distribution with t-distribution which more accurately estimate population parameter.Also, t-distribution is useful for estimating population parameter even if it is not normallay distributed.For these reason,I can tell that assumptions of the test used in question 4 met.

- The End