Jiankun_Dong_CS555_HW2

Jiankun (Bob) Dong CM3226

2023-10-06

Problem 1:

```
kid_cal <- read.csv("kid_cal.csv")
Kid_meal <- kid_cal$Calories[kid_cal$trt == T]
Kid_no_meal <- kid_cal$Calories[kid_cal$trt == F]

IQparticipant <- 1.5*(quantile(Kid_meal,.75)[[1]]-quantile(Kid_meal,.25)[[1]])
outlierMeal <- Kid_meal < quantile(Kid_meal,.25)[[1]]-IQparticipant | Kid_meal > quantile(Kid_me
al, .75)[[1]]+IQparticipant
#sum(outlierMeal) There are no outliers
mealFrame <- data.frame(
  Mean = mean(Kid_meal),
  Median = median(Kid_meal),
  SD = sd(Kid_meal),
  First_Quantile = quantile(Kid_meal,.25)[[1]],
  Third_Quantile = quantile(Kid_meal,.75)[[1]],
  Min = min(Kid_meal),
  Max = max(Kid_meal),
  outlier = "NULL"
)

IQNonparticipant <- 1.5*(quantile(Kid_no_meal,.75)[[1]]-quantile(Kid_no_meal,.25)[[1]])
outlierNoMeal <- Kid_no_meal< quantile(Kid_no_meal,.25)[[1]]-IQNonparticipant | Kid_no_meal > qu
antile(Kid_no_meal,.75)[[1]]+IQNonparticipant

noMealFrame <- data.frame(
  Mean = mean(Kid_no_meal),
  Median = median(Kid_no_meal),
  SD = sd(Kid_no_meal),
  First_Quantile = quantile(Kid_no_meal,.25)[[1]],
  Third_Quantile = quantile(Kid_no_meal,.75)[[1]],
  Min = min(Kid_no_meal),
  Max = max(Kid_no_meal),
  outliner = Kid_no_meal[outlierNoMeal]
)
```

Summary of Calorie for Participants of Meal Preparation:

```
(mealTable <- kable(mealFrame,"simple"))
```

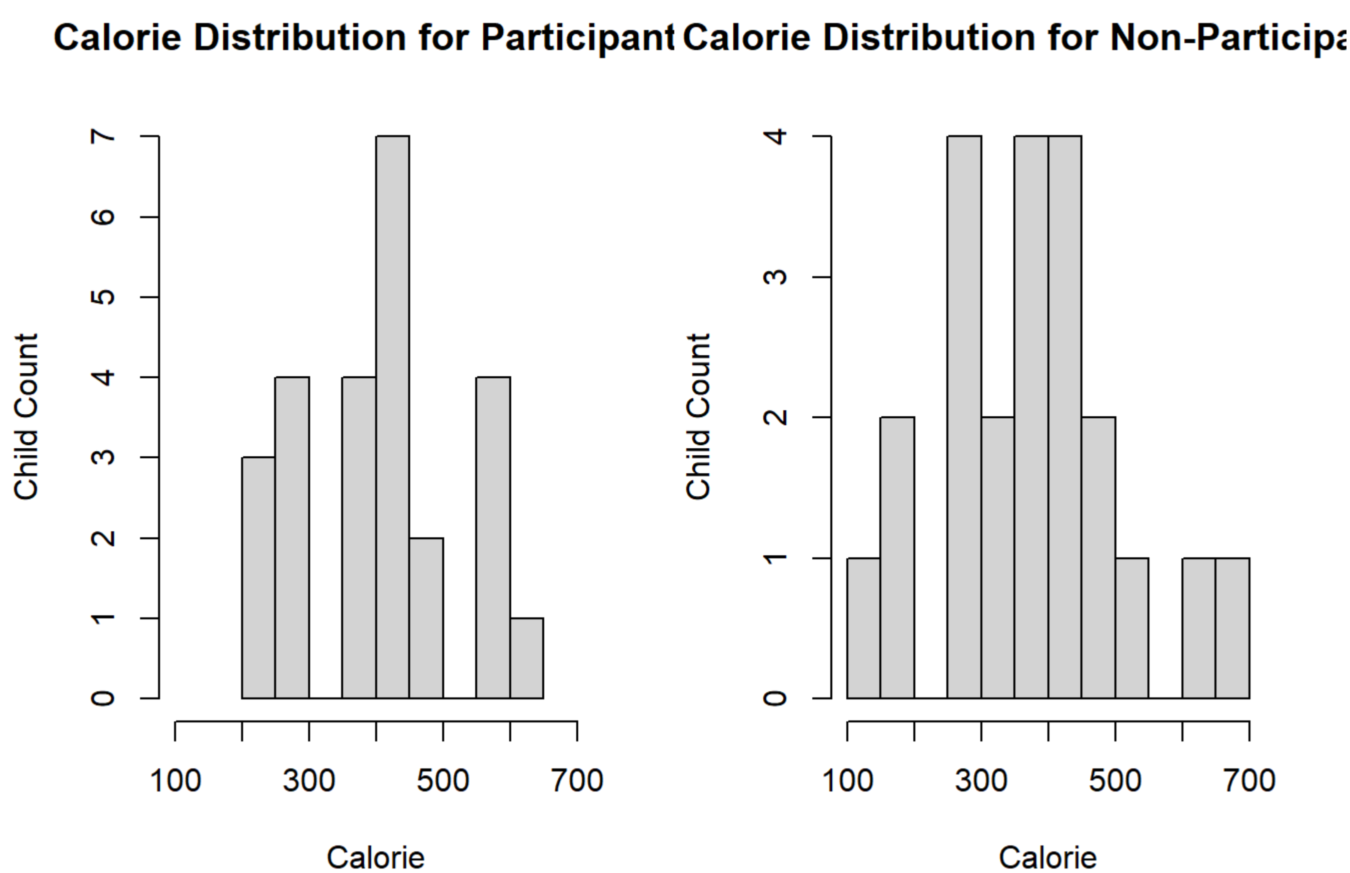
Mean	Median	SD	First_Quantile	Third_Quantile	Min	Max	outlier
410.0796	424.94	121.5138	298.38	456.3	210.99	635.21	NULL

Summary of Calorie for Non-Participants of Meal Preparation:

```
(nomealTable <- kable(noMealFrame,"simple"))
```

Mean	Median	SD	First_Quantile	Third_Quantile	Min	Max	outliner
374.0718	374.74	133.1393	296.3925	445.5575	139.69	688.77	688.77

```
par(mfrow = c(1,2))
hist(Kid_meal, main = "Calorie Distribution for Participants",
     xlab = "Calorie", ylab = "Child Count",breaks = 10, xlim = c(100,700))
hist(Kid_no_meal, main = "Calorie Distribution for Non-Participants",
     xlab = "Calorie", ylab = "Child Count",breaks = 10, xlim=c(100,700))
```



The graph of non-participant's meal calorie distribution follows roughly a normal distribution, with ONE outlier on the right side of the graph.

The graph of participant's meal calorie distribution also roughly follows a normal distribution, but without any outlier.

Both graph have similar shape, but non-participant's graph has a wider range of calorie.

Problem 2:

```
alpha <- 0.05
n <- length(Kid_meal)
```

Step 1: $H_0 : \mu 0 = 425$ $H_1 : \mu 1 \neq 425$ $\alpha = 0.05$ and $n = 25$, df = 24
Step 2: because the population σ is unknown and the sample is small, use the t test where $t = \frac{\overline{x}-\mu}{\frac{S}{\sqrt{n}}}$
Step 3: Decision rule: Reject H_0 if $t \geq 2.064$ or $t \leq -2.064$
Step 4:

```
xbar <- mean(Kid_meal)
S <- sd(Kid_meal)
t <- (xbar - 425)*sqrt(25)/S
```

the t value is -0.6139386
Step 5: Do not reject H_0
We do not have strong evidence that at confidence level $\alpha = 0.05$ that the mean calorie consumption for those who participated in the meal preparation differ from 425.

Problem 3:

```
meal_test <- t.test(Kid_meal,conf.level = .9)
```

The 90% confidence interval's lower bound is 368.5004482, and upper bound is 451.6587518.
This means that at 90% confidence level, we will reject hypothesis that states the mean of the participant's meal calorie is below368.5004482, or higher than 451.6587518.

Problem 4:
Step1: $H_0 : \mu 1 = \mu 2, H1 : \mu 1 \geq \mu 2, \alpha = 0.05, df = 21$ where non-participant correspond to $\mu 2$
Step2: $t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu 1 - \mu 2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$
Step3: Decision rule: we reject H_0 if $t \geq 1.721$
Step4:

```
t_5 <- (mean(Kid_meal)-mean(Kid_no_meal))/sqrt(sd(Kid_meal)^2/25+sd(Kid_no_meal)^2/22)
```

Step5: Fail to reject H_0 because 0.9636039 is not greater than 1.721
We do not have significant evidence that at $\alpha = 0.05$ level that participants consume more calorie than non-participants.

Problem 5:
The is indeed one outlier in the non-participant dataset. Furthermore, we don't know the method the data was taken, therefore we can't say with confidence that the samples are independent.