CS555, Data Analysis and Visualization Homework 2

Jefferson Parker, [japarker@bu.edu](mailto:japarker@bu.edu)

See the accompanying R code at the end of this document.

1. For the two groups of children, Participants in the Meal Preparation arm of the study vs Non-Participants, the distributions of calorie intake were different in terms of their numeric values and visual inspection of their distributions.



Participant children took in about 100 more calories than non-participant children comparing any of the minimum, median, or maximum measurements. This difference is visually apparent in both box plots and histograms (common x-scale) of the data, below.



Furthermore, the non-participant histogram is suggestive of having some left-skewness, with some values at the lower end of the range. The participant histogram shape is very rough, with adjacent calorie ranges having more or fewer children and could probably benefit from having fewer cuts in the plot.





1. To test if mean calorie consumption for meal preparation participants is different from 425 (calories) with  = 0.05:
   1. Our hypotheses are:
      1. H0:  = 425
      2. Ha:  > 425
   2. We want to calculate t as our test statistic, as our sample size (n) is less than 30 and we are not provided a population standard deviation.
      1. The critical value for  = 0.05 with df = 24 is t = 1.71.
      2. Our decision rule is to reject H0 if t >= 1.71.
   3. The test statistic is calculated to be t = 0.30.

Given that our test value is not greater than our critical value, we cannot reject the null hypothesis.

1. The 90% confidence interval for the mean calorie intake for participants in the meal preparation arm of the study is 428.7 - 434.1 calories. This result indicates that the population mean of children following the meal preparation plan has a 90% likelihood of being in this range.
2. To test if meal preparation participants actually consume more calories than non-participants with with  = 0.05:
   1. Our hypotheses are:
      1. H0:  = 2
      2. Ha:  > 2
   2. We want to calculate t as our test statistic, as our sample size (n) is less than 30 and we are not provided a population standard deviation.
   3. For this situation, we will test for a p-value less than our critical value. Our decision rule is reject H0 if p < .
   4. The test statistic is calculated to be p = 0.00352, so we reject H0 and accept the alternative.
3. The assumptions for test four are:
   1. The samples are randomly selected and independent. We must trust that this is true since we are not given enough information to determine otherwise.
   2. The calorie intake of both sample groups was measured by the same method. Again, we must assume this is also true given that we have no information to the contrary.
   3. The data should be normally distributed, or at least similarly shaped and without outliers. From the result in part 1, we know that the data is similarly shaped and when tested there were no outliers detected.

Given the above stated caveats, the assumptions of the two-sample t-test were met.

**R Code:**

# CS555 Data Analysis and Visualization

# Homework2.R

# Jefferson Parker, japarker@bu.edu

# 20180713

# Load the calorie intake data.

inputDir <- "C:/Users/jparker/Code/Input";

setwd(inputDir);

inCalories <- read.table(file = "CalorieIntakeForparticipants.txt", header = TRUE);

outCalories <- read.table(file = "CalorieIntakeForNon-participants.txt", header = TRUE);

# Add some labeling and annotation.

colnames(inCalories) <- "calories";

inCalories$Group <- rep("In", length(inCalories$calories));

colnames(outCalories) <- "calories";

outCalories$Group <- rep("Out", length(outCalories$calories));

calData <- rbind(inCalories, outCalories);

# 1. Summarize the data by participation status.

# Use an appropriately labeled table and visualization.

# Describe the distribution shape and comment on similarities.

summary(inCalories);

summary(outCalories);

boxplot(calories ~ Group, data = calData, main = "Calorie Intake by Study Participant Class");

hist(inCalories$calories, xlim = c(100,700), main = "Calorie Intake, Participants", xlab = "Calories");

hist(outCalories$calories, xlim = c(100,700), main = "Calorie Intake, Non-Participants", xlab = "Calories");

# 2. For participants, test if the mean is different (two sided) from 425 with alpha = 0.05.

# Use the t-test because n < 30 and we do not know population standard deviation.

qt(0.05, length(inCalories$calories) -1, lower.tail = FALSE);

t.test(inCalories$calories, mu = 425, alternative = 'two.sided', conf.level = 0.05);

# 3. Calculate and interpret the 90% confidence interval for participants.

# The confidence interval is independent of the test value, so we don't need to compare to any mean.

# 90% confidence level requires alpha = 0.1.

# OR If calculating manually to make sure the t.test with mu = 0, use P = 0.45 (half the area).

t.test(inCalories$calories, alternative = 'two.sided', conf.level = 0.1);

xbar <- mean(inCalories$calories);

se.xbar <- sd(inCalories$calories)/sqrt(length(inCalories$calories));

t.crit <- qt(0.45, length(inCalories$calories)-1);

ci.upper <- xbar + (t.crit\*se.xbar);

ci.lower <- xbar - (t.crit\*se.xbar);

ci.lower;

ci.upper;

# 4. Test whether participants consumed MORE than non-participants at alpha = 0.05.

t.test(x = inCalories$calories, y = outCalories$calories, alternative = 'greater', conf.level = 0.05);

# 5. Test the data for the presence of outliers.

find.outliers <- function(x){

fiveNum <- fivenum(x);

firstQuartile <- fiveNum[2];

thirdQuartile <- fiveNum[4];

iqrVal <- fiveNum[4] - fiveNum[2];

upperBound <- thirdQuartile + (1.5 \* iqrVal);

lowerBound <- firstQuartile - (1.5 \* iqrVal);

return(sort(x[x <= lowerBound | x >= upperBound]));

}

find.outliers(outCalories$calories);

find.outliers(inCalories$calories);