Riley Payung

STAT 250-003

Data Analysis Assignment #4

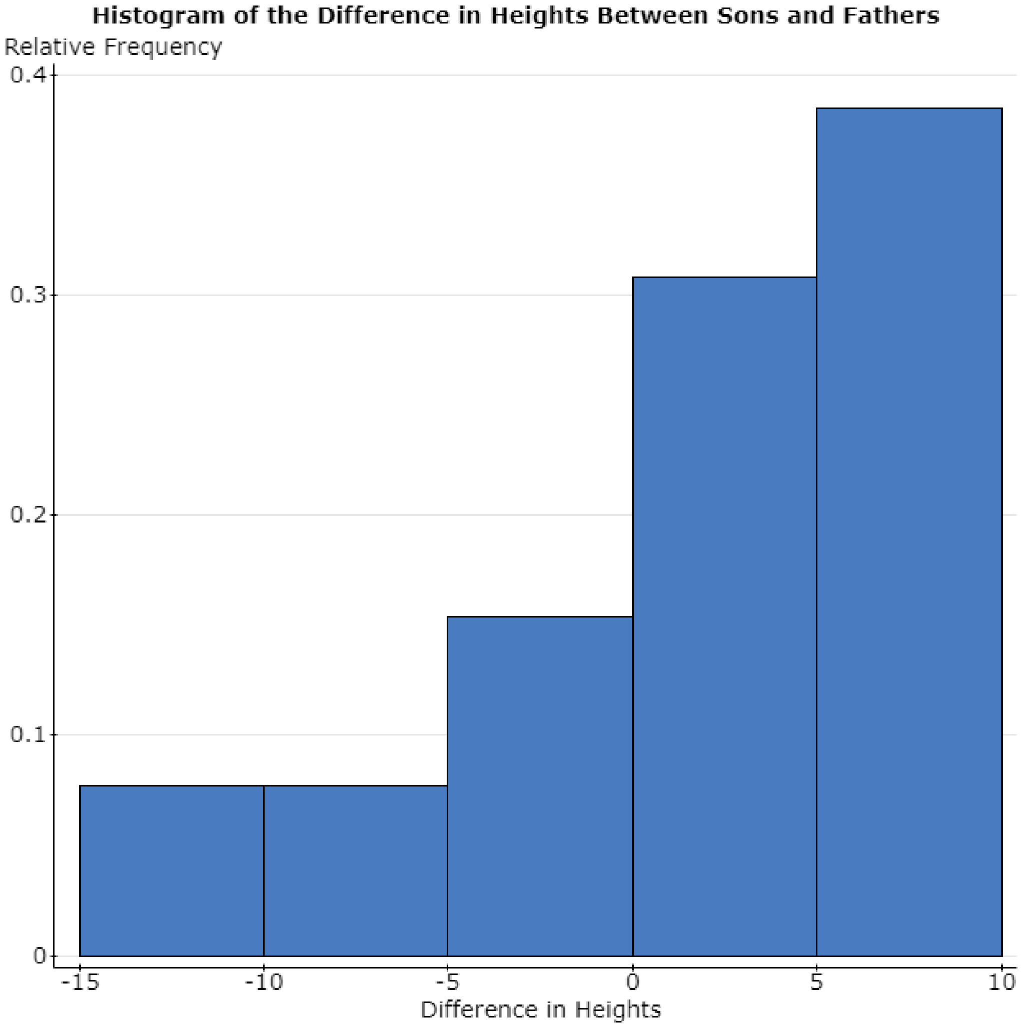
Problem #1

Heights of Fathers and Sons. To test the claim that sons are taller than their fathers on average, a researcher randomly selected 13 fathers who have adult male children. She records the height of both the father and son in inches

1A

μ_D is our parameter of interest; we want to find the differences in the two heights and use those differences to deduce a conclusion.

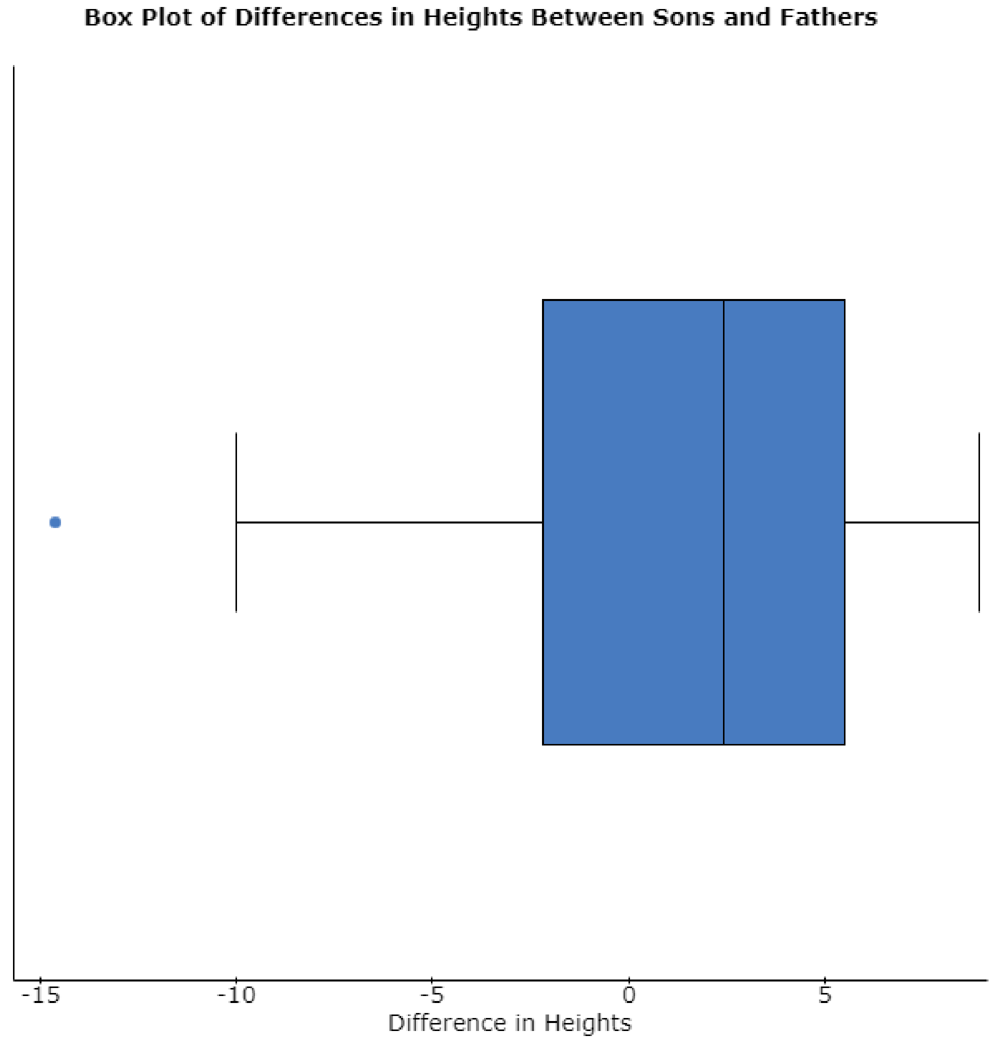
1B



1C

The shape of the histogram is left-skewed.

1D



1E

There is a single outlier at -14.60

1F

Statistical Inference would be inappropriate here because we have outliers. Really, we only have one outlier, but it will still skew the data heavily, as we can see in the histogram. The boxplot shows us that we have an outlier.

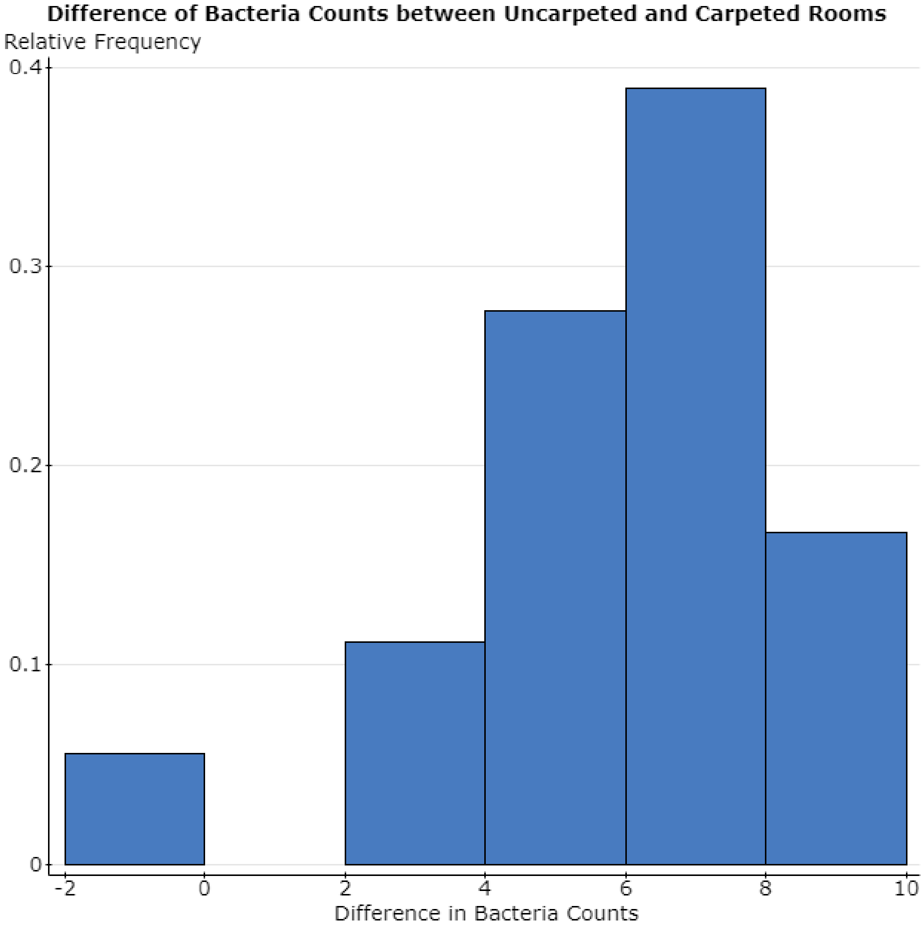
Problem #2

Bacteria Counts. Researchers wanted to determine if carpeted rooms contained more bacteria than uncarpeted rooms. To determine the amount of bacteria in a room, researchers pumped the air from the room over a Petri dish at the rate of 1 cubic foot per minute for a random selection of eighteen carpeted rooms and another random sample of eighteen uncarpeted rooms in a very large hospital. Colonies of bacteria were allowed to form in the Petri dishes. The data are presented as the count of bacteria per cubic foot.

2A

μ_D is our parameter of interest; we want to find the differences in the bacteria counts between the two rooms and use those differences to deduce a conclusion.

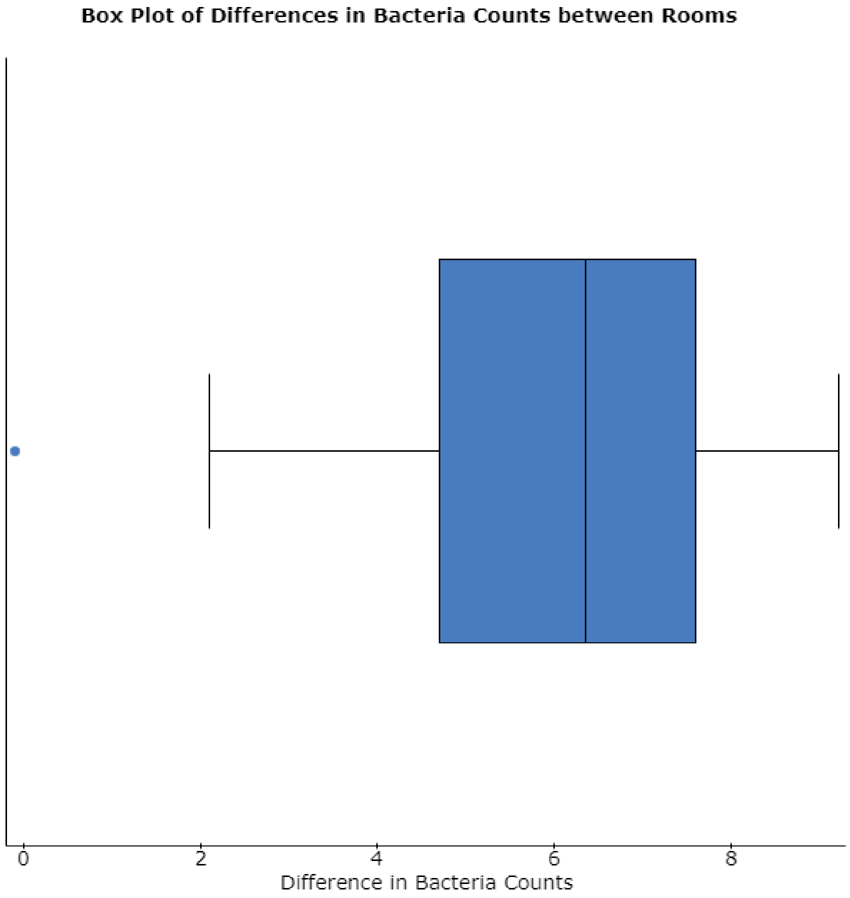
2B



2C

The shape is minorly skewed to the left.

2D



2E

We have an outlier at -0.1.

2F

Statistical Inference would not be appropriate in this case because we have an outlier. Our N is small enough (< 25). Based on the box plot, we have an outlier, and our histogram would be relatively normal if we removed said outlier.

Problem #3

Researchers randomly selected participants to take part in a study. The participants were randomly assigned either a tall, thin “highball” glass or a short, wide “tumbler,” each which held 355 ml. The participants were asked to pour a shot (1.5 oz. = 44.3 ml) of water into their glass. Did the shape of the glass make a difference in how much liquid they poured? Assume all conditions for conducting inference are satisfied. Use the following sample statistics to complete this problem.

3A

We want to know if the type of glass they used changed the amount of water they poured; we’re looking for the difference in the amount of water they poured.

3B

Two sample T summary confidence interval:

μ_1: Mean of Population 1
μ_2: Mean of Population 2
$\mu_1 - \mu_2$: Difference between two means
(without pooled variances)

95% confidence interval results:

Difference	Sample Diff.	Std. Err.	DF	L. Limit	U. Limit
$\mu_1 - \mu_2$	18.7	2.4263911	194.08012	13.91452	23.48548

3C

H_0: $\mu_D = 0$

H_a: μ_D != 0 (Not Equal)

3D

Since the **confidence interval does not capture 0**, which we are testing for in our null hypothesis, we can conclude that there is a significant difference in the amount of water that the individual poured, based on the summary statistics.

3E

Two sample T summary hypothesis test:

μ_1: Mean of Population 1
μ_2: Mean of Population 2
$\mu_1 - \mu_2$: Difference between two means
H_0: $\mu_1 - \mu_2 = 0$
H_A: $\mu_1 - \mu_2 \neq 0$
(without pooled variances)

Hypothesis test results:

Difference	Sample Diff.	Std. Err.	DF	T-Stat	P-value
$\mu_1 - \mu_2$	18.7	2.4263911	194.08012	7.7069192	<0.0001

I can verify that we reject the null hypothesis since the p-value is <0.0001.

Problem #4

A particular county's Health Department experimented with a flexible four-day workweek. For a year, the department recorded the mileage driven by 11 field workers on an ordinary five-day workweek. Then, it changed to a flexible four-day workweek and recorded the mileage for another year for the same 11 field workers. Test the hypothesis that the five-day workweek has a greater average mileage. Assume all conditions are satisfied in this problem. The data set used for this problem is called "Flexible Work Schedule." Use a significance level of 0.10.

4A

The population parameter in this case in the difference in mileage driven in 1 year for the same 11 workers in the Health Department.

4B

H_0: $\mu_D = 0$

H_a: $\mu_D > 0$

4C

Employee Name	5 Year	4 Year	Difference
Jeff	2798	2914	-116
Betty	7724	6112	1612
Roger	7505	6177	1328
Ali	838	1102	-264
Sera	4592	3281	1311
Claire	8107	4997	3110
Lida	1228	1695	-467
Perez	8718	6606	2112
Enzo	1097	1063	34
Greg	8089	6392	1697
Martin	3807	3362	445

4D

Summary statistics:

Column	n	Mean	Std. dev.
Difference	11	982	1139.5683

4E

Test Statistic: 982 – 0 / (1139.5683 / sqrt(11)) = 2.858

4F

Paired T hypothesis test:

$\mu_D = \mu_1 - \mu_2$: Mean of the difference between 5 Day and 4 Day
H_0: $\mu_D = 0$
H_A: $\mu_D > 0$

Hypothesis test results:

Difference	Mean	Std. Err.	DF	T-Stat	P-value
5 Day - 4 Day	982	343.59278	10	2.8580344	0.0085

4G

We reject the null hypothesis since our p-value is less than alpha at 0.0085 < 0.10.

4H

Since we have significant evidence to reject the null hypothesis, we can conclude that the recorded mileage of a five-day workweek has a greater average mileage than the four-day workweek.

Problem #5

A poll randomly surveyed 1508 US residents aged 13 – 65 asking about their sleep habits, and, in particular, their use of technology around the time they try to go to sleep. Research shows that people who regularly use their computers in the hour before trying to go to sleep are less likely to report getting a good night's sleep. The poll found that of the 19-29 years old sampled, 205 out of 293 reported using a computer in the hour before trying to go to sleep. In contrast, of the 30 – 45-year olds sampled, 313 out of 469 reported computers use an hour before trying to sleep.

5A

The population parameter is Americans getting a good night's sleep after using their technology for an hour.

5B

n is most definitely greater than 25, and the population size is at least 10x bigger than the sample (300 million).

Big Population: 15080 < ~300 million. n = 1508, bigger than 25.

5C

Sample Proportions:

p-hat$_1$ = 205 / 293 = 0.6996.

p-hat$_2$ = 313 / 469 = 0.6674.

Parameter Estimate:

p_D = 0.6996 – 0.6674 = 0.0322

5D

Two sample proportion summary confidence interval:

p_1: proportion of successes for population 1
p_2: proportion of successes for population 2
$p_1 - p_2$: Difference in proportions

99% confidence interval results:

Difference	Count1	Total1	Count2	Total2	Sample Diff.	Std. Err.	L. Limit	U. Limit
$p_1 - p_2$	205	293	313	496	0.06861031	0.03444701	-0.0201193	0.1573399

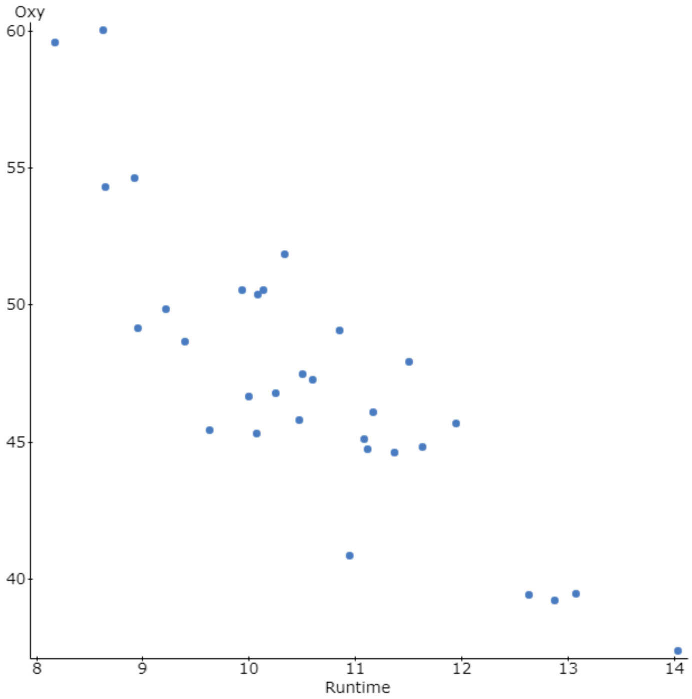
5E

Yes, the interval captures 0, this means that we accept the null hypothesis.

Problem #6

Data from a physical fitness program was collected on 31 men that were asked to run 1.5 miles. Variables measured include oxygen uptake in ml/min (Oxy), their resting pulse rate before the run in beats per minute (RstPulse), their time to run 1.5 miles in minutes (Runtime), the pulse rate at the end of the run in beats per minute (RunPulse), their maximum pulse rate during the run in beats per minute (MaxPulse), their weight in kg (Weight), and their age in years (Age). The StatCrunch data is called Physical Fitness. Investigate the relationship between the explanatory variable “Runtime” and response variable “Oxy” by doing the following:

6A



6B

R (correlation coefficient) = -0.86215152

6C

I would say that the trend of the graph is strongly negative, since our R value is close to -1, and has a slightly curved shape.

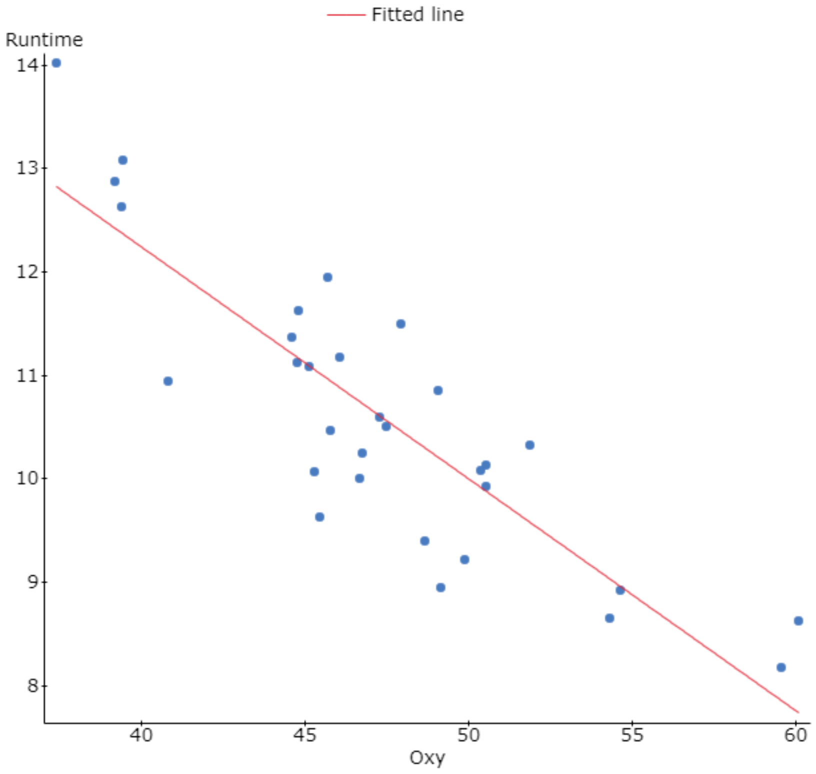
6D

Simple linear regression results:

Dependent Variable: Runtime
Independent Variable: Oxy
Runtime = 21.22219 - 0.22450253 Oxy
Sample size: 31
R (correlation coefficient) = -0.86215152

R-sq = 0.74330524
Estimate of error standard deviation: 0.71495093

6E



6F

Y = 21.22219 - 0.22450253x

6G

Theoretically, the average pulse rate when running would be 21.22219 with the y-intercept of 21.22219.

6H

Its important because it tells us the slowest person in this case and gives us a base point to attach the regression to.

6I

R-sq = 0.74330524

The model of this regression line fits the time in which it takes to run the 1.5 miles best when the R-squared value is closest to 1; seeing as it is 0.743, it is a generally good model to fit the situation.

6J

21.22219 - 0.22450253(17) = 17.40564699

Based on the model that fits the data, this individual would complete the run using 17.4 oxy, and would be way off the graph.

6K

Yes, because its not even on the graph in this case.