STA305/1004 in-class problem - Solutions

Nathan Taback

2017-02-14

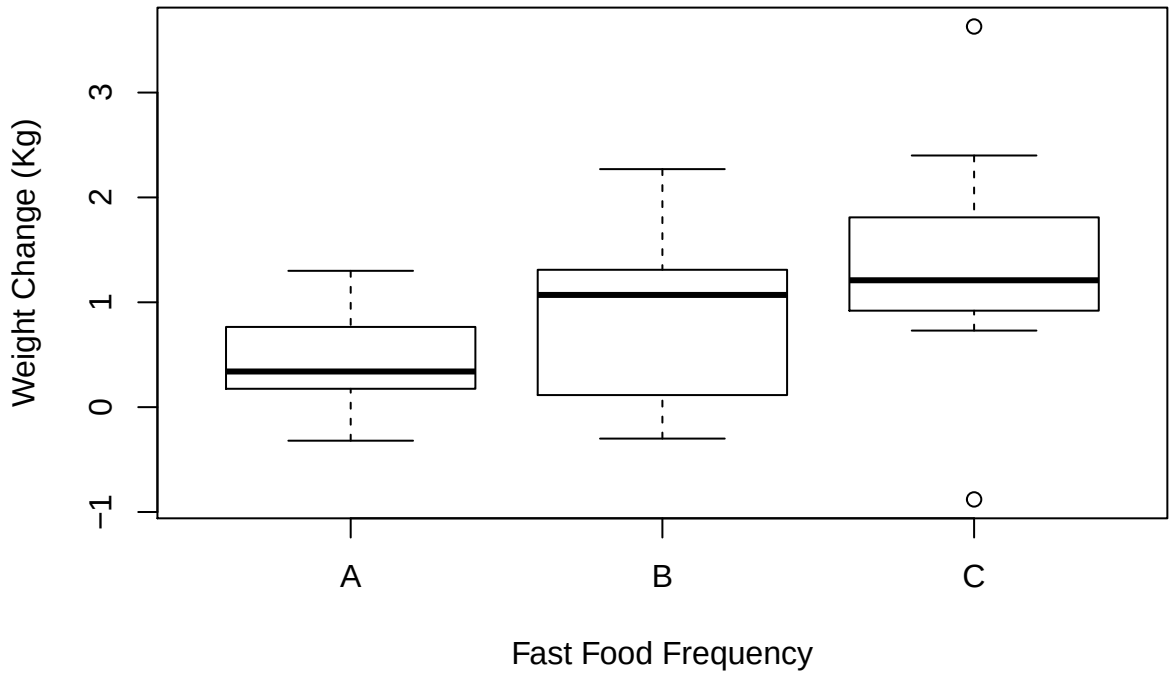
Answer the questions based on the following experimental design.

A study at UofT recruited twenty-one students to complete a thirty minute survey on their diet and eating habits at the end of an academic year. Students were paid $10 to complete the survey and answer a few questions. The data below shows their weight gain from September to April classified by the frequency that students ate fast food. In group A students reported eating fast food once per month; the students in group B reported eating fast food twice per month; and the students in group C reported eating fast food four times per month.

	A	B	C
	1.02	1.44	0.73
	-0.32	0.40	1.11
	0.27	-0.30	3.63
	0.08	2.27	-0.88
	0.51	-0.17	1.21
	0.34	1.07	1.22
	1.30	1.18	2.40
Treatment Average	0.46	0.84	1.35
Treatment SD	0.55	0.92	1.40

The researchers analyzed the data using R.

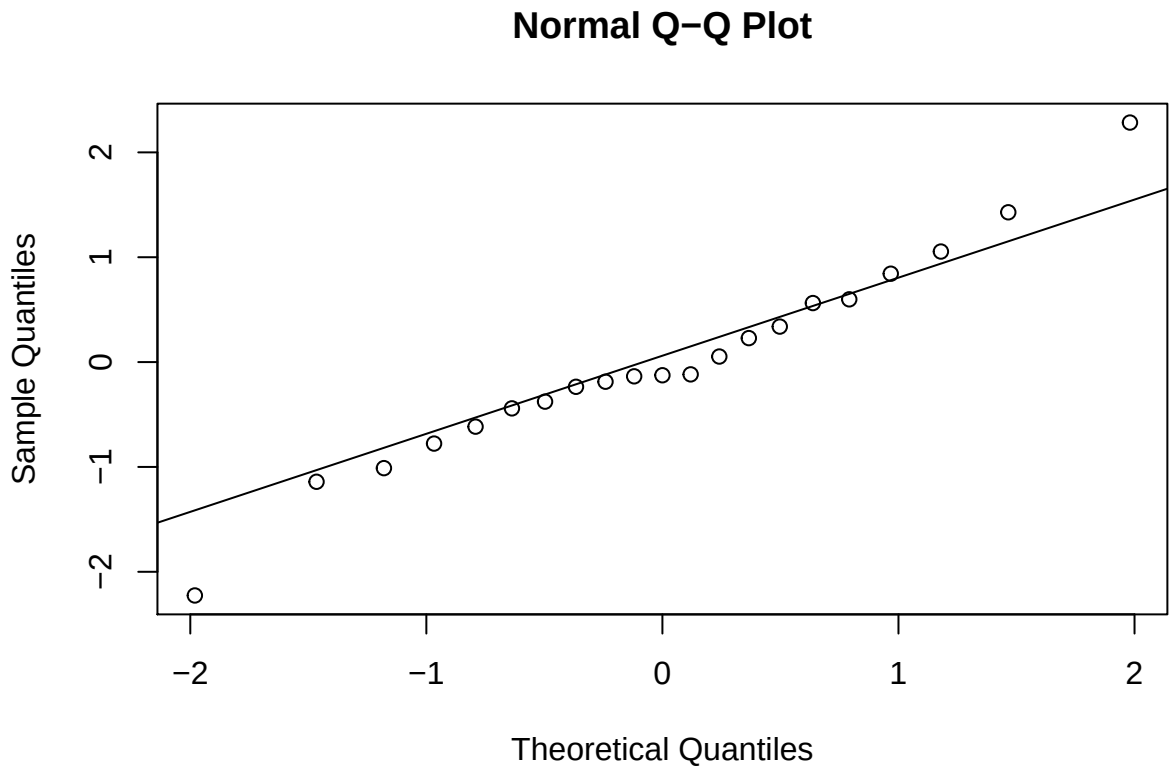
```
surveydat <- read.csv("surveydat.csv")
boxplot(wtchange~grp,data = surveydat,ylab="Weight Change (Kg)", xlab="Fast Food Frequency")
```



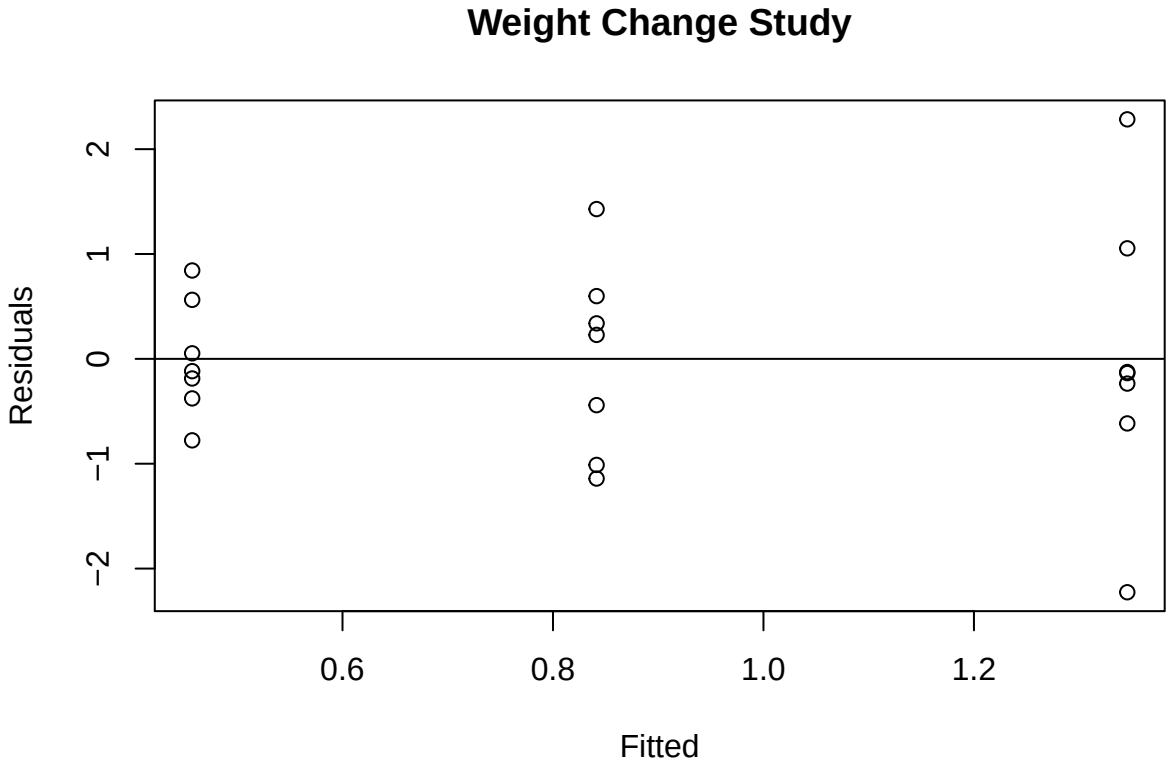
```r
aovsurvey <- aov(wtchange~grp,data=surveydat)
summary(aovsurvey)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## grp          2   2.78   1.390   1.341  0.287
## Residuals   18  18.66   1.037
```

```r
qqnorm(aovsurvey$residuals);qqline(aovsurvey$residuals)
```




```r
plot(aovsurvey$fitted.values, aovsurvey$residuals,ylab="Residuals",
     xlab="Fitted",main="Weight Change Study")
abline(h=0)
```



Questions

(a) Is this study and experiment or observational study? What are the treatments?

This is an observational study. The treatment has three levels: eating fast food once per month; eating fast food twice per month; and eating fast food four times per month. Treatment assignment is non-ignorable since

(b) Would it have been feasible for the researcher to randomized students to the treatments? What randomization scheme (assigning the subjects to the treatments) could the researcher use to accomplish the randomization?

A randomized study is not feasible since the treatment is the frequency of eating fast food. Nevertheless, a hypothetical randomization scheme could be constructed by labeling all the students using the numbers 1 to 21 then obtaining a random permutation of these numbers. Assign the first 7 students to diet A, the next 7 to diet B, etc.

(c) What are the null and alternative hypotheses that the researchers are testing in the data analysis? Is there evidence to reject the null hypothesis? Explain.

$$H_0 : \mu_1 = \mu_2 = \mu_3 \text{ vs. } H_1 : \mu_i \neq \mu_j, i \neq j.$$

There is no evidence to reject H_0 since the p-value is 0.287.

(d) Are the statistical assumptions behind the data analysis satisfied? Explain.

The three assumptions are outlined below.

1. Additive model.

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}.$$

This seems plausible since change in weight from each diet can be viewed as the sum of a common mean plus a random error term.

2. The errors ϵ_{ij} are independent and identically distributed (iid) with common variance $Var(\epsilon_{ij}) = \sigma^2$, for all i, j

- The common variance assumption can be investigated by plotting the residuals versus the fitted values of the ANOVA model.

- A plot of the residuals versus fitted values can be used to investigate the assumption that the residuals are randomly distributed and have constant variance.
- In this case the points don't fall randomly on both sides of 0. The residuals are increasing as the fitted values increase. This is an indication that the common variance assumption is not satisfied. This can also be seen from the standard deviations in each treatment group: the largest (1.4) is approximately three times as large as the smallest (0.55).
- We are not given any information to confirm that that observations are independent. For example, if some of the students in the sample roommates then their weight gains and fast food consumption may not be independent.

3. $\epsilon_{ij} \sim N(0, \sigma^2)$.

- The normal quantile plots indicates that this assumption is satisfied since the points fall along the straight line.

It is not the case that all the assumptions are satisfied since the non-constant variance asssumption may not be true. In addition, it's difficult to confirm if the data are independent. Therefore, it might be the case that the p-value is not accurate (e.g., the p-value might actually be smaller.)

(e) The researcher is convinced that the results of the study would have provided strong evidence that eating fast food four times per month causes students to gain weight, if the sample size in each group was larger. Is this a valid statement? Explain.

This is not a valid statement. Consider the following points:

- There is no comparison group where students did not eat fast food so it's not possible to calculate the causal effect of fast food versus no fast food on weight gain. It is possible to calculate the non-causal effect of eating less fast food versus more fast food.
- Subjects assigned the "treatment" to themselves so we don't know if there are differences in the types of students (e.g., age, sex, history of being overweight) that selected themselves to be in the groups.
- If the sample size in each group was larger then the power would increase. Although, even if the study was designed to have high power and the analysis yielded a small p-value then this still wouldn't fix the way treatment was assigned. So, we wouldn't know if the differences are due to the treatment or due to differences between the types of students in the groups.