STAT 423 Project Report

Daniel Yang, Peter Bol, Max Zhang, Miki Wan

Introduction

This sampling design project is intended to investigate students' food budget on campus. Various questions surrounding students' habit with food will In this manner, students can make informed decisions on how to budget food based on their needs. Our survey is designed to a mixture of nominal and ratio questions, which allows us to discuss results both qualitatively and quantitatively. Population parameters to consider are population totals, means, proportions, and variances. Additionally, sample size, costs, and confidence interval calculations will also be handled.

Survey Design and Data Collection

The survey is delivered electronically through classlists of 4 different introductory University courses in different faculties. We chose the 4 introductory courses because of the large class size and diversity of students from different faculties. Our goal is to capture as many students from different faculties as possible.

The survey contains a few identifier questions such as "Gender", "Age", and "Student's faculty" to help us understand the sample population. To investigate students' food budget and what factors influence students' food budget, questions about students' habits around food are asked. We believe these questions may make a difference in students' budget for food.

Questions used for the investigations are listed below: How much do you usually spend on food on campus per week? ($_________ per week) Which meal(s) do you eat on campus most frequently? (select up to 3) What are your preferred cuisine(s)? (Select up to 3) On average, how many caffeinated beverages do you consume per week? Do you track calories, macronutrients, or other quantifiable food measures?

To prevent non-responses with individual questions, we designed the survey as such that every question is mandatory before a survey taker submits the survey. Since we are sampling from 4 different classes, we aimed to sample around 10 students from each class. The classes varied in class sizes with the smallest being 57, and the largest being 401. We received 42 samples in total.

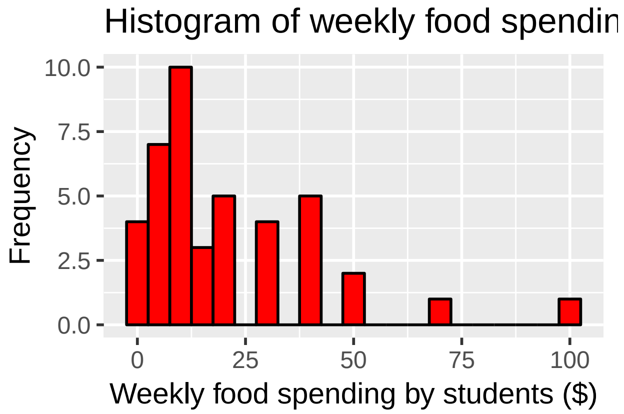
Sampling Method

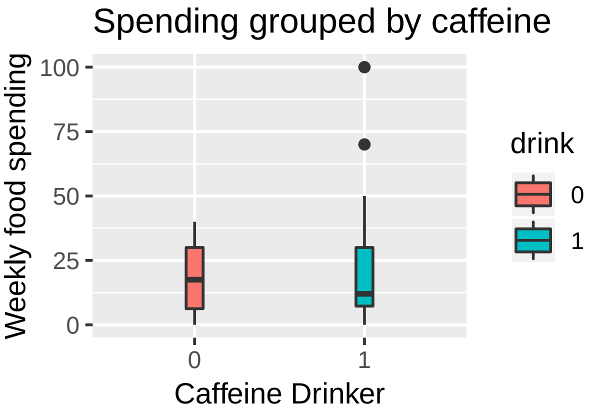
We use the four different as our clusters for the whole school population. In a two stage cluster sample, we first choose our clusters, then we sample a subpopulation within each clusters. Our sampling frame consist of a list of student ID from each of the four classes. The total number of winter courses (i.e., the primary sampling units) is $N = 1435$, and the number of winter courses we sampled from is $n = 4$. For the secondary sampling units, the grand total number of students (undergraduates and graduates) from the University of Calgary factbook is $\sum_{i=1}^{1435} M_i = 32436$, but since all students can take several courses (thereby violating the disjoint cluster assumption), we cannot compute $\overline{M}$, instead we estimate the quantity with $\overline{m}$. The classes chosen were general electives in order to increase the variability in the kind of students sampled, and although attempts were made to control cluster overlapping (for example, conflicting schedules), by the nature of probability sampling there will still be overlapping issues.

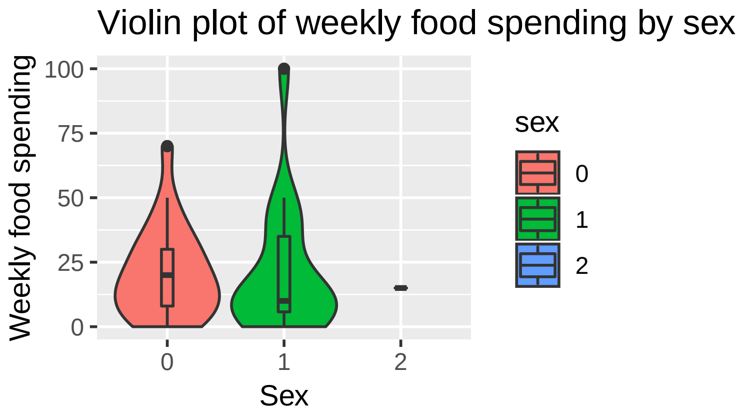
Analysis

Descriptive Statistics

We begin our investigation by first looking at the descriptive statistics and plots of our dataset for some combination of our variables, in order to identify possible (or the lack thereof) relationships. Among the many plots that are looked at for this purpose, below are 3 that merits investigation.







The histogram indicates the data on weekly food spending is right skewed. Interestingly enough, it appears that there isn't really a difference in weekly food spending by people that do not drink caffeine beverages at all versus people that drink caffeine at least once. Furthermore, there appears to be some disparity in how weekly food spending occurs between the sexes - based on the quartiles and the violin plot, it seems that females are more variable in their spending than males (there is only one observation unit marked as 'fluid', so no inference can be made for that specific factor). Of course, these graphical checks should be supplemented with formal tests. Below are the summary statistics for our main variable of interest, weekly food spending, by cluster:

```
##   class min    Q1 median   Q3 max     mean       sd  n missing
## 1     0   0  5.75   10.5 27.5  40 16.80000 14.92053 10       0
## 2     1   5 12.50   20.0 30.0  70 25.45455 19.80588 11       0
## 3     2   0 10.50   22.5 30.0 100 27.00000 28.84441 10       0
## 4     3   0  5.00   12.0 12.5  50 14.00000 16.18641 11       0
```

Inferential Statistics

It is befitting to begin the analysis with the most common statistics, the mean and total, along with their standard errors and 95% confidence intervals, for weekly food spending. The manner of calculation for those quantities is the usual manner for 2-stage cluster samples, at that was the sample design. Below is a compiled table of our results:

Table 1: Estimated population mean and total for average weekly food spending, along with standard errors and 95% confidence intervals.

Statistic	Estimate	Std. Error	95% CI
$\hat{\mu}$	21.69	3.81831	[13.96, 29.42]
$\hat{T}$	6443835	1134210	[4147747, 8739923]

We find that the average of average weekly food spending to be 21.69 dollars, and the total of average weekly food spending to be 6443835 dollars (a future follow up to this survey might be to actually verify the amount of revenue food vendors receive on a weekly basis). Considering the nature of how our data is distributed (in particular, the skewness), it might be of some interest to us to find the proportion of students who do not spend any money on food on campus at all. This estimation would allow us to qualify in some sense how "meaningful" the mean statistic is - if the proportion of students who do not spend any money on food is large, it may be worthwhile to look into the median instead, as a more robust statistic for our purposes.

Table 2: Estimate population proportion of students who do not spend any money on food on campus, along with standard errors and 95% confidence intervals.

Statistic	Estimate	Std. Error	95% CI
$\hat{p}$	0.1119	0.02198	[0.06882 0.15496]

The proportion of students who do not purchase any food is estimated to be about 11.19%. While not exactly a small proportion, it's not so large that a mean estimation is irrelevant. One interesting thing to note is how *low* our computed standard errors are compared to the standard errors within each cluster. This disparity arises primarily with how large our $\bar{m}$ is, as the classes we chose to sample from generally have a much larger student enrollment. As a result, we are underestimating our 2-stage cluster variances by a substantial degree, so some care should be taken prior to placing a faith in this undeserved precision. We should also be mindful of the actual sample size required to make such estimations; for example, to have an error of *just 25%*, the sample size required for the mean is

$$\begin{aligned} n &= \frac{N \cdot S_R^2}{N \cdot \left(\frac{e\bar{m}}{z_{\alpha/2}}\right)^2 + S_R^2} \\ &= 1024.368 \end{aligned}$$

about 1025 samples, of which the formula had to be altered, as the variance had to be estimated (and as discussed, likely underinflated). It is possible to reduce the number of samples by selecting a different sampling design, but in any case it is clear that more samples are required for even a conservative estimate.

Since the sampling design is a 2-stage cluster, we hope to see heterogeneity within each cluster, and homogeneity between each cluster. Typically, this is checked by computing the intra-cluster correlation coefficient (ICC) and checking the sign, but the ICC requires the cluster sizes to be all the same. Since clearly class sizes can deviate, an adjusted R squared calculation was made instead, with some quantities obtained via ANOVA table below:

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## class         3   1291   430.4    1.02  0.395
## Residuals    38  16034   422.0
```

We compute $R^2_{\text{Adj}} = 1 - \frac{422}{S^2} = 0.001363$. This low value is not surprising, since our sampled primary sampling units are quite high, and intuitively food spending for a student should not be expected to be correlated to the course they are enrolled in. Interpretation of variability in this case suggests that variation within and between clusters are about the same. An added bonus from conducting the ANOVA is it tests for equal means in the 4 sampled clusters. The p-value is reported to be 0.395; at $\alpha = 0.05$ there is not enough evidence against the null hypothesis to claim the means of each cluster are different. However, this interpretation is contingent on the assumption that the sampled data is from a Gaussian distribution, which was not verified in this project.

Relative efficiency was also computed, with respect to a simple random sample design. We find

$$\text{RE}(\hat{\mu}, \hat{\mu}_{\text{SRS}}) = \frac{\text{Var}(Y)/n \cdot (1 - \frac{n}{N})}{\text{Var}(\hat{\mu})} = 7.23$$

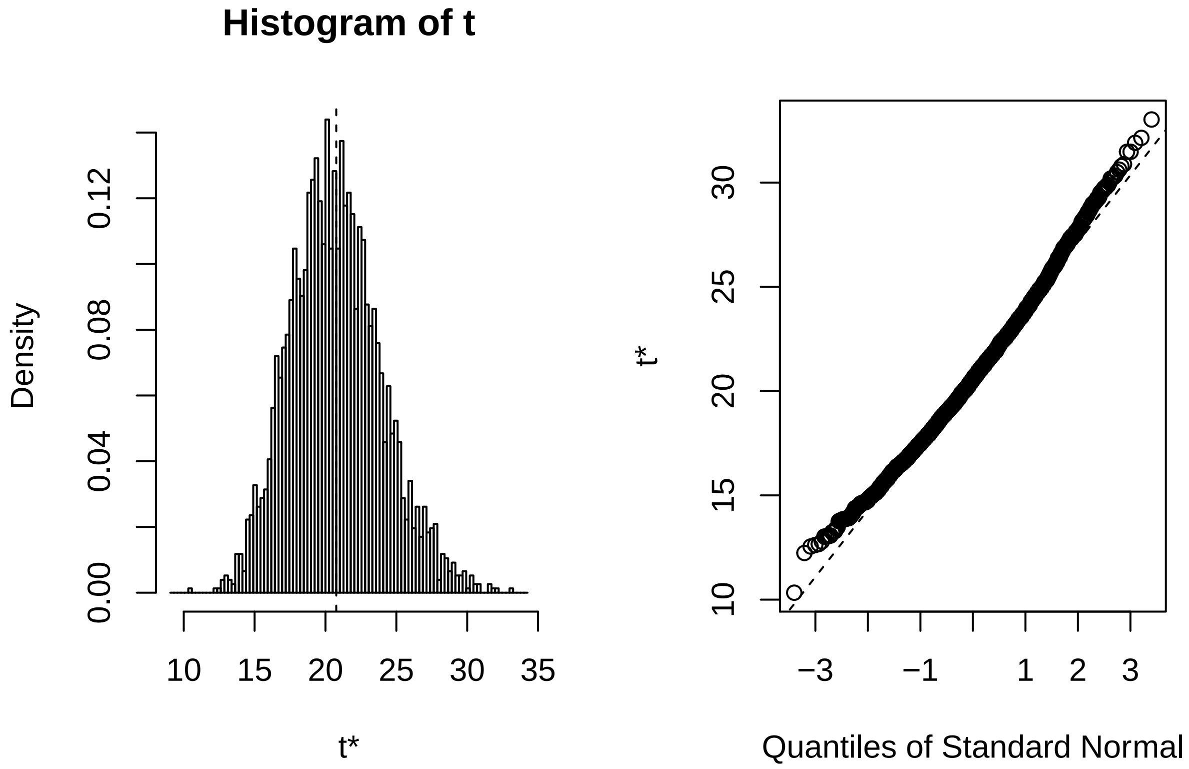
At a glance this result says that the cluster sampling is more efficient than simple random sampling (SRS) from a variance point of view. This result is worth commentating because it is misleading. In general, cluster sampling is known to be not as efficient of a design as simple random sampling or stratified sampling. The cluster variance has been mentioned to be underestimated, and the results of the R^2_{Adj} indicate the design of both cluster sampling and SRS should perform similarly well. Finally, even under extreme conditions where cluster sampling is significantly better than an SRS, a relative efficiency of 7.23 is staggering. Despite all this, this result is still included in the report to illustrate the dangers of blindly relying on statistics without context.

Bootstrapping

Given the fact that the sample size obtained in this study is so small, it would be remiss to not consider the bootstrap to improve estimation (and, while not covered in this project, offers a suitable alternative to hypothesis testing when parametric assumption are not met - the t-test, for example). A few bootstrap statistics and their distributions are considered, with a few being used in later analyses to demonstrate their uses and applications. Below is the bootstrap distribution for the mean of the average weekly food spending by students:

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = dfstat$dt, statistic = boot.mean, R = 3000)
##
##
## Bootstrap Statistics :
##     original      bias    std. error
## t1*  20.7619 -0.03830952    3.211341
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 3000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = tmean, type = "perc")
##
## Intervals :
## Level     Percentile
## 95%   (14.91, 27.64 )
## Calculations and Intervals on Original Scale
```

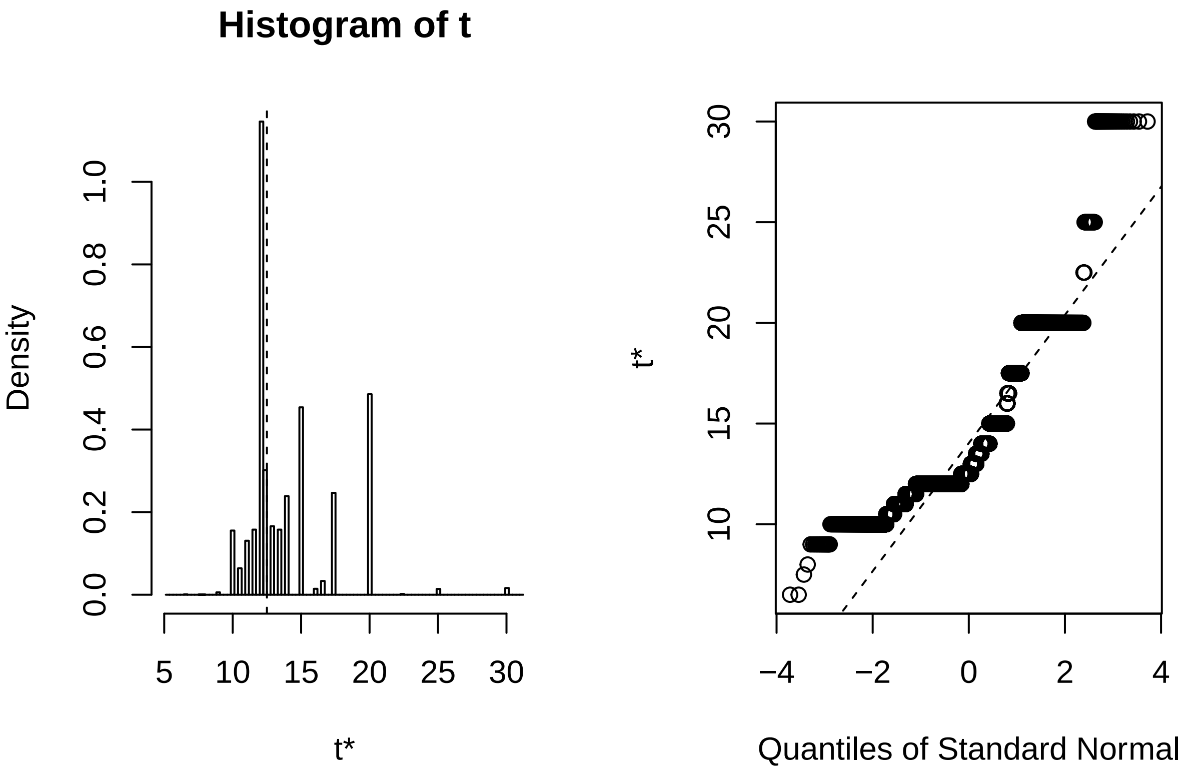


Not too surprisingly, the bootstrap statistic has a similar value in both the estimate and the standard error to the previously calculate sample mean, at 20.76 and 3.21 respectively. The 95% confidence interval based on the percentile approach evaluates to [14.91, 27.64]. Despite the sample data being heavily skewed, the distribution of the bootstrap resembles that of a normal distribution - this was entirely expected, as per the results of the Central Limit Theorem, which states the sampling distribution of the mean asymptotically approaches a normal distribution with mean μ and variance σ^2/n. A replication of 3000 was considered. Next, we consider the bootstrap distribution of the median of average weekly food spending.

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = dfstat$dt, statistic = boot.median, R = 10000)
##
##
## Bootstrap Statistics :
##     original  bias    std. error
## t1*     12.5 1.53025    3.180562
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = tmed, type = "perc")
```

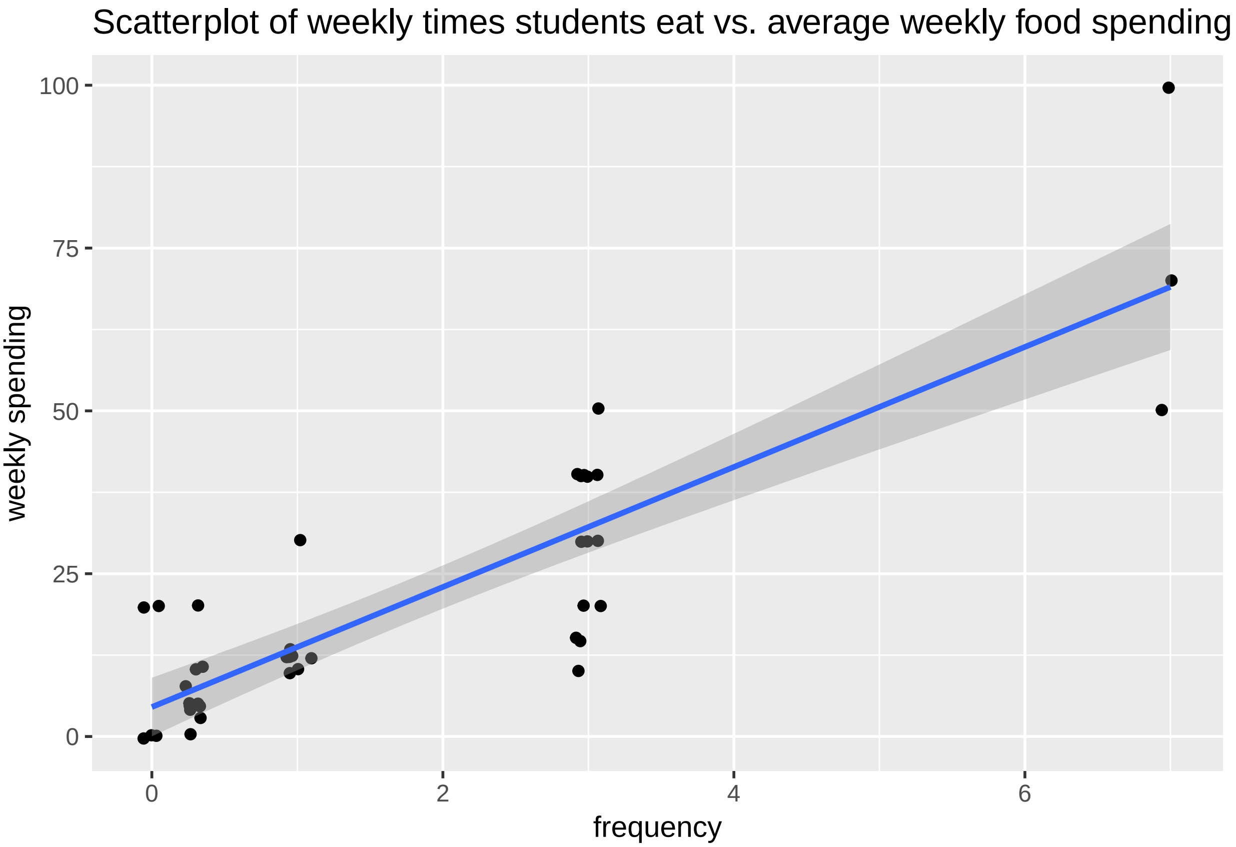
```
##
## Intervals :
## Level     Percentile
## 95%   (10, 20 )
## Calculations and Intervals on Original Scale
```

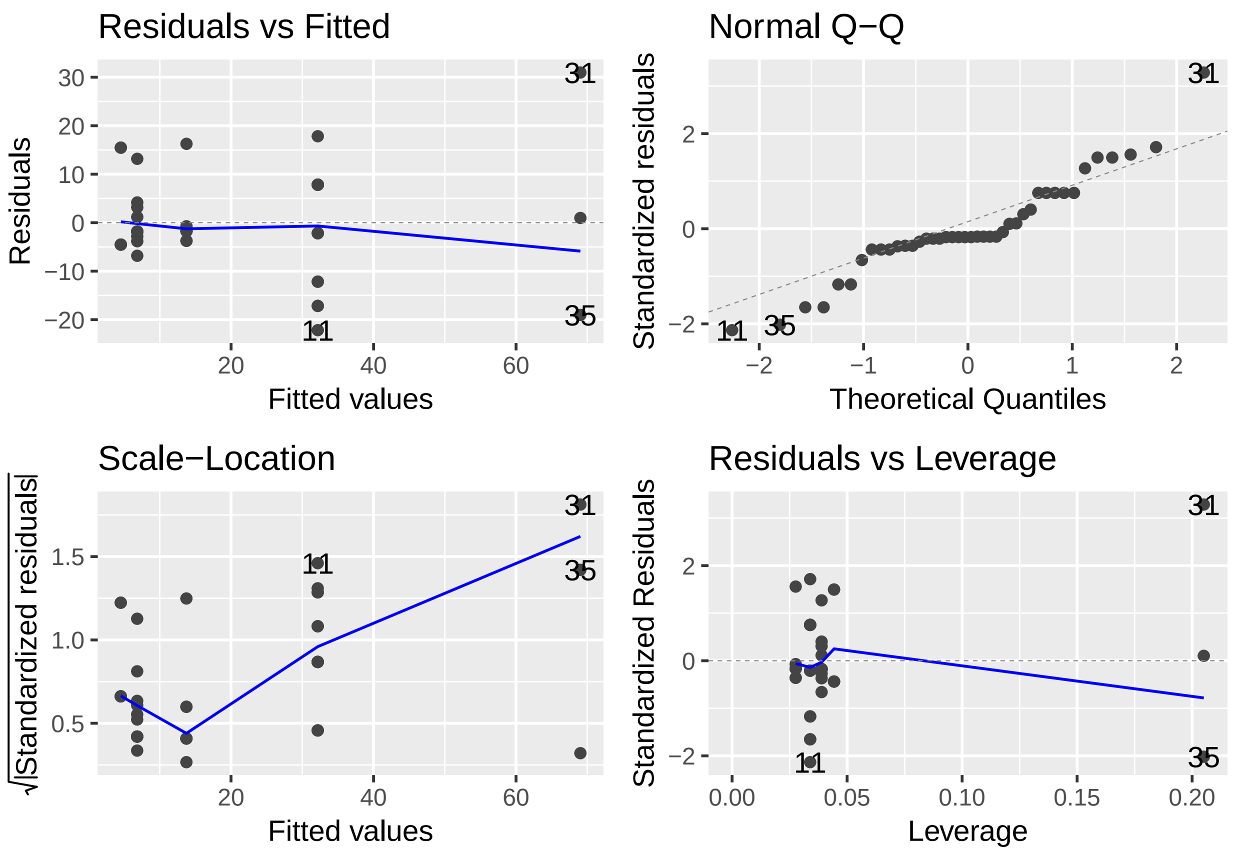


The bootstrapped median was found to be 12.5, with a standard error of 3.18. Despite running 10000 replicates (in contrast to 3000 for the mean), the distribution is still skewed - the Central Limit Theorem says nothing for the median, after all.

Ratio and Regression Estimation

In an attempt to further improve estimation of average weekly food spending, we consider ratio and regression estimation. An appropriate auxiliary variable to consider is the frequency of times students buy food on campus, since there should be a clear positive association. The responses to this question was standardized to a timeframe reference of a week (i.e., responses like ‘once a month’ is encoded instead as 0.25, ‘every day’ is 7, etc.). Below is a scatterplot of the two variables, and some diagnostics plots to ensure assumptions for regression estimation is satisfied.


Scatterplot of weekly times students eat vs. average weekly food spending
weekly spending
100
75
50
25
0
0
2
4
6
frequency



There is indeed a positive correlation with frequency of consumption and spending. The diagnostics plots also indicates adequacy - there is no pattern in the residuals plot, and while the errors do not seem normal this should improve as sample size increases. Both ratio and regression estimation requires knowing the population parameter of the auxiliary variable (in this case, the population mean of times students buy food on campus); since we do not know that quantity, we estimate it via the bootstrap statistic.

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = Survey$foodfreq, statistic = boot.mean, R = 3000)
##
##
## Bootstrap Statistics :
##     original        bias    std. error
## t1* 1.761905 9.920635e-05   0.2868212
```

The bootstrap estimate $\bar{X}_{\text{freq}} = 1.76$; the estimated average amount of times per week students buy food on campus is 1.76. How reliable is this estimate requires us to know how well the sampled data 'resembles' the population - in this case, our sample size is likely too small to say with certainty (although a value of 1.76 is certainly *credible*, if not likely). Regardless, we proceed with our analysis using this statistic. Below is a table that summarizes our results.

method	estimate	std. error
Ratio	20.76	3.47
Regression	20.76	5.28

Both methods produced the same estimates for average weekly food spending (20.76), however ratio estimation has a lower standard error. This is surprising, since ratio estimation should only fare better when the intercept is 0, and from the graph above that does not appear to be the case. A regression model was fitted to check against this:

```
##
## Call:
## lm(formula = weekly ~ foodfreq, data = Survey)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.171  -3.808  -1.830   6.914  30.969
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.5259     2.2259   2.033   0.0487 *
## foodfreq      9.2150     0.8594  10.723 2.49e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.57 on 40 degrees of freedom
## Multiple R-squared:  0.7419, Adjusted R-squared:  0.7354
## F-statistic:   115 on 1 and 40 DF,  p-value: 2.486e-13
```

Unfortunately, it does seem like the intercept is marginally significant at $\alpha = 0.05$. However, recall that the slope and intercept coefficients are themselves an estimate based on the sampled data. We can try to simulate repeated sampling via bootstrapping to verify if these results are indeed statistically significant (if they are, they should hold in the bootstrap as well). Below are the results of the bootstrap:

```
##    Names           Coefs              SEs
## X1 "(Intercept)"   "4.52591427447897" "4.20624327353172"
## X2 "foodfrequency" "9.21502162799843" "1.62424648329544"
##    zVals              Pvals
## X1 "1.07599917079424" "0.281927627247928"
## X2 "5.67341331674122" "1.39979947110217e-08"
```

The results from the bootstrap indicate that the intercept $\hat{\beta_0}$ is not statistically significant at $\alpha = 0.05$, and so we fail to reject $H_0 : \beta_0 = 0$, which would explain why ratio estimation has a lower standard error than regression estimation. It is also fortunate that the slope β_1 is still found to be significant.

Hypothesis Testing

We considered our choice of hypotheses based on interesting graphical relationships found earlier and expectations going into this project.

Hypothesis 1.

We test for equal variances between the male and female groups for average weekly spending. The violin plot yielded an interesting distribution for the female groups and a more volatile quantile, although both groups are skewed in the same manner. Our hypothesis test is

$$H_0 : \sigma^2_{\text{Male}} = \sigma^2_{\text{Female}}$$
$$H_1 : \sigma^2_{\text{Male}} \neq \sigma^2_{\text{Female}}$$

We conduct a Levene's test. Below is the output:

```
##
##  modified robust Brown-Forsythe Levene-type test based on the
##  absolute deviations from the median
##
## data:  df$weekly
## Test Statistic = 0.39567, p-value = 0.6759
```

at $\alpha = 0.05$ we fail the reject the null hypothesis; there is not enough statistical evidence present in the sample that the variation in the two groups are any different.

Hypothesis 2.

We test if the means of weekly food spending for participants that drink caffeinated beverage vs. those that do not are different. Since coffee, tea, etc. are generally incorporated into a food budget we would have expected a difference (the boxplots indicate otherwise, at least for the median). Our hypothesis test is

$$H_0 : \mu_{\text{Caffeine Drinkers}} = \mu_{\text{Non-caffeine drinkers}}$$
$$H_1 : \mu_{\text{Caffeine Drinkers}} \neq \mu_{\text{Non-caffeine drinkers}}$$

We conduct a two sample t-test. Below is the output:

```
##
##  Welch Two Sample t-test
##
## data:  weekly by drink
## t = -0.55233, df = 36.541, p-value = 0.5841
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -15.344352   8.772923
## sample estimates:
## mean in group 0 mean in group 1
##        18.57143        21.85714
```

at $\alpha = 0.05$ we fail to reject the null hypothesis; there is not enough statistical evidence against the null hypothesis that the means of the two groups are equal.

Hypothesis 3.

At the beginning of the data analysis we expected to see food preferences and type of meals eaten to matter to this project in some sense, such as forming a hypothesis or being a usable auxiliary variable. This did not turn out to be the case. Below are 2 outputs after fitting into a regression model:

```
##
## Call:
## lm(formula = weekly ~ asia + medi + amer + mexi, data = Survey)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.723 -11.734  -5.066   9.314  63.150
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   23.314      7.658   3.044  0.00428 **
## asia          -5.128      7.198  -0.712  0.48070
## medi          13.536      7.311   1.851  0.07209 .
## amer          -6.453      6.708  -0.962  0.34236
## mexi          -3.905     10.471  -0.373  0.71131
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.16 on 37 degrees of freedom
## Multiple R-squared:  0.1321, Adjusted R-squared:  0.03831
## F-statistic: 1.408 on 4 and 37 DF,  p-value: 0.2503
```

```
##
## Call:
## lm(formula = weekly ~ breakfast + lunch + dinner, data = Survey)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.143 -11.831  -3.111   6.310  63.857
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.111      6.173   0.504  0.61722
## breakfast     12.926      7.032   1.838  0.07386 .
## lunch         20.106      6.461   3.112  0.00352 **
## dinner         8.187      8.911   0.919  0.36402
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.67 on 38 degrees of freedom
## Multiple R-squared:  0.2357, Adjusted R-squared:  0.1753
## F-statistic: 3.906 on 3 and 38 DF,  p-value: 0.01586
```

As observed, food preference of any kind is not statistically significant, and only lunch is significant. In other words, there are no linear associations (aside from lunch) from this sample between all the mentioned covariates and average weekly food spending.

Discussion

Upon analyzing the responses from our survey, we found that several things could be improved to increase the quality of our data and the clarity of the questions to survey takers. These improvements are listed below:

1. The responses to frequency questions should be proportional so that we when we assign a number to each response for analysis purposes, it is more representative to the proportionality of the frequency. Instead of "Never, Once a month, Once a week, Few times a week, always", we could ask for the frequency per week, or month. This will allow us to conduct our analysis with a set unit.

2. In the question that asks the survey taker to input the amount of money they spend on food on campus, we should specify whether or not beverages are included in this budget. Many would consider a drink with a meal as part of their spending on food.

3. We found that the response rate for people's daily calorie intake/ macronutrient level given that they track these measurements is very low. This may be caused by the ambiguity of the response format. Instead of short answers for the macronutrient level, we could implement a fill in the blank with each of the three macronutrients listed clearly. These non-responses are missing at random, since it's possible to estimate such quantities as these responses are correlated in some way (e.g., we would expect higher calorie count to be correlated with larger average weekly spending on food).

These are methods of improvement that we could implement and keep in mind in prospective sample design.

Clear questions and responses will make it easier for the survey takers and provide better data for our analysis. Another improvement we could make to improve our sampling design is to take the samples from classes in proportion to the class sizes. This will give us an even sampling weight between our clusters. Finally, it is pertinent to mention that while the design of the question is important to eliminating bias, the relevancy of the questions themselves must also be considered. For future discussion, more relevant questions such as how long one stays on campus, the income of the observation unit, or their living proximity to campus should be asked as well for stronger associations and better predictions.

Conclusion

We find the average weekly spending on food by the sampled students to be 21 dollars and 69 cents. There are no statistical difference in variability in spending between males and females and no statistical difference in the mean in spending between coffee drinkers and non coffee drinkers. There is no association between food preference or type of meal consumed (saved for lunch) and amount of weekly spending. Ratio estimation produces the lowest standard error. As the emphasis was placed on analysis of results rather than the results themselves, it would be more appropriate to characterize this report as a parable rather than a story. But ironically enough, not a lot of precision is required to find the average weekly expenditure, since there are only so many credible values to take in the first place.