[redacted]

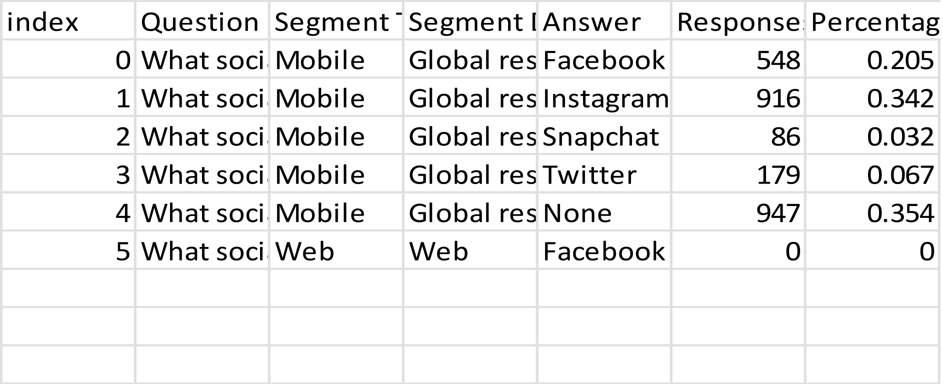
1/16/2023

Intro to Stats (MATH – 1223-01)

Dr. Jacob

Project 1 Dataset

My dataset can be found at https://www.kaggle.com/datasets/thedevastator/uncovering-millennials-shopping-habits-and-socia and below this paragraph. Adam Halper is the main source for this data set. My dataset's title is the "Social Influence on Shopping" and includes a social survey of data from 300,000 millennials and gen z members. The survey that is given to these participants determines how they make their decisions when it comes to shopping, and which social media plat form influences their purchases the most. The two quantitative variables are the count column and the percentage column. The two categorical variables are the segment type and segment description.



index	Question	Segment	Segment I	Answer	Response	Percentag
0	What soci	Mobile	Global res	Facebook	548	0.205
1	What soci	Mobile	Global res	Instagram	916	0.342
2	What soci	Mobile	Global res	Snapchat	86	0.032
3	What soci	Mobile	Global res	Twitter	179	0.067
4	What soci	Mobile	Global res	None	947	0.354
5	What soci	Web	Web	Facebook	0	0

The variables in this study are both the 300,000 millennials and the gen z members, the survey that the participants were given and the answers in which they chose. One reason I chose this study was because I am a member of Gen Z and I find this topic interesting because I can relate to this study. I think it's interesting that they used a university for the major source of their participants. Over 60% of the feedback came from a university, which I am also in. I relate to this data set because I am in Gen Z and in college and want to know more about the influence the social world has on spending, especially for college students.

I hope to find out how gen z and millennials are pressured and influenced to spend money. I find myself easily influenced through social media and peer pressure to buy things that are trendy, so I hope to see that I am not alone in my purchased being influenced by other people.

Answer Labels	Total Responses	Relative Frequency
Facebook	1582	0.210120866
Instagram	2560	0.340018595
None	2659	0.353167751
Snapchat	239	0.031743923
Twitter	489	0.064948864
Grand Total	**7529**	**1**

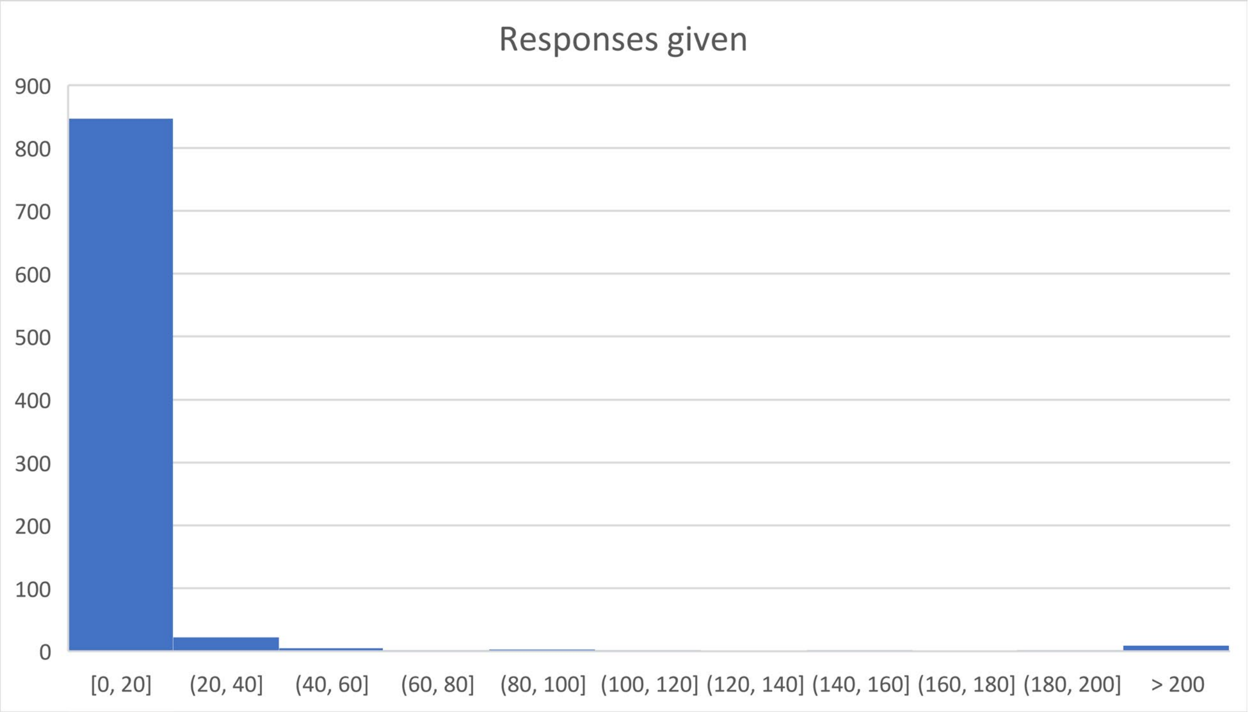
This is the sum and relative frequency of the answers that the respondents gave. This is the grand total of actual respondents in all categories.

Row Labels	Facebook	Instagram	None	Snapchat	Twitter	Grand Total
Gender	2	2	2	2	2	10
Mobile	1	1	1	1	1	5
University	174	174	174	174	174	870
Web	1	1	1	1	1	5
Grand Total	**178**	**178**	**178**	**178**	**178**	**890**

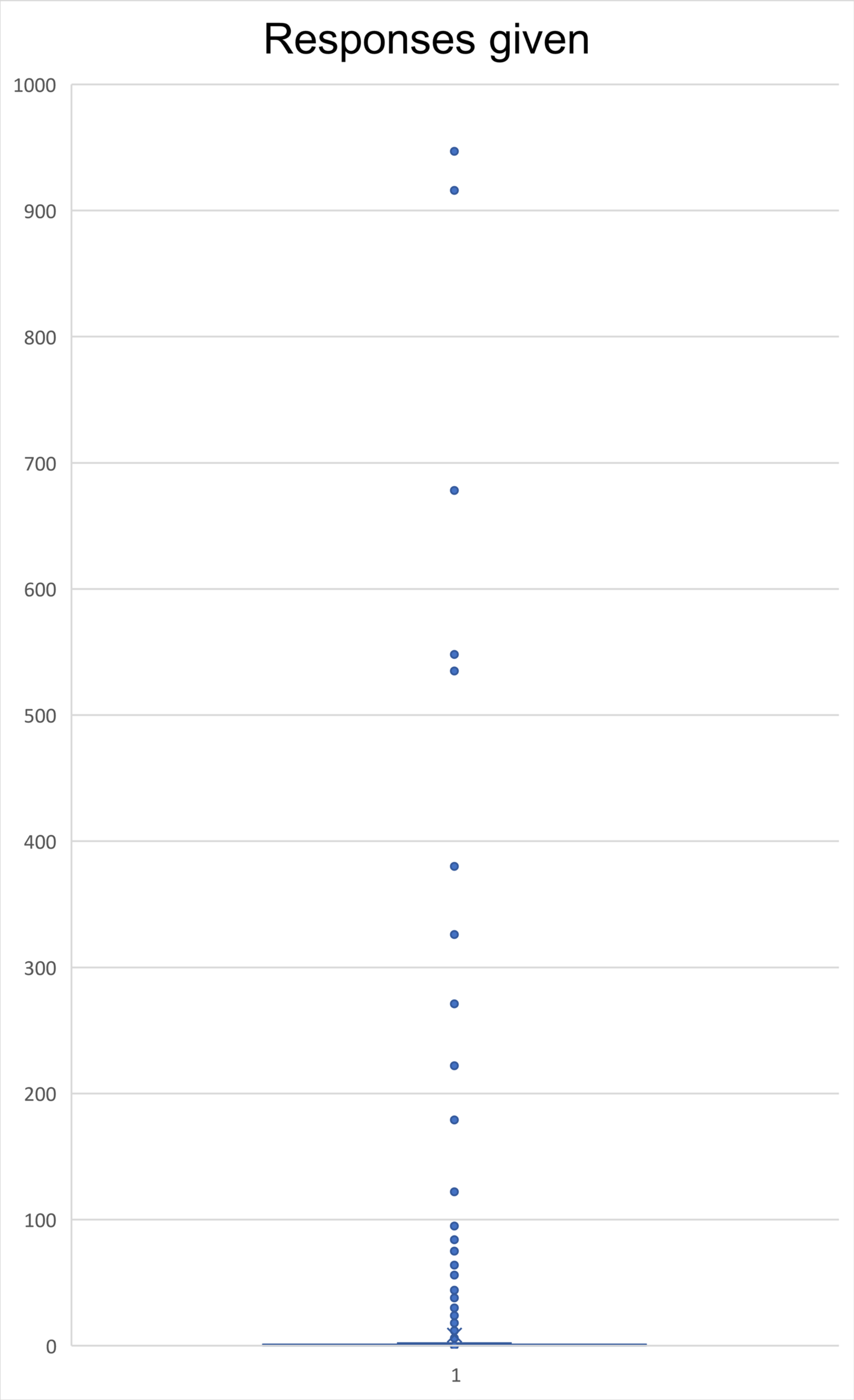
Above is the segment types surveyed with answer that was given for the survey. The numbers for each method are proportional because all rows, even ones with no value are currently included in the table. With the snippets of data shown, at this time I have not found a conclusive relationship between the two variables. The grand total corresponds with the number of rows in the data set.

The quantitative variable used is the number of responses from the survey. Below is the mean, standard deviation and five number summary of the variable.

Mean:	8.459550562
Std Dev:	60.10881625
Min:	0
Q1:	0
Q2:	0
Q3:	1
Max:	947



Above is a histogram showing how many total responses grouped by twenties, that each row had with an overflow of two hundred for an ease of reading. The distribution is right skewed.



Above is a box plot of the total of responses. There are several outliers. The range makes it difficult to see the box because it spans a large group of numbers and the outliers skew the chart.

For our quantitative hypothesis, we'll be looking at the average of how many responses were given (different than the number of rows).

H_0: $\mu = 10$
H_1: $\mu \neq 10$

For our categorical hypothesis, I am testing if social media has influenced purchasing habits. The parameter that I am looking at is the answers. We will be analyzing if the p score is statistically significant.

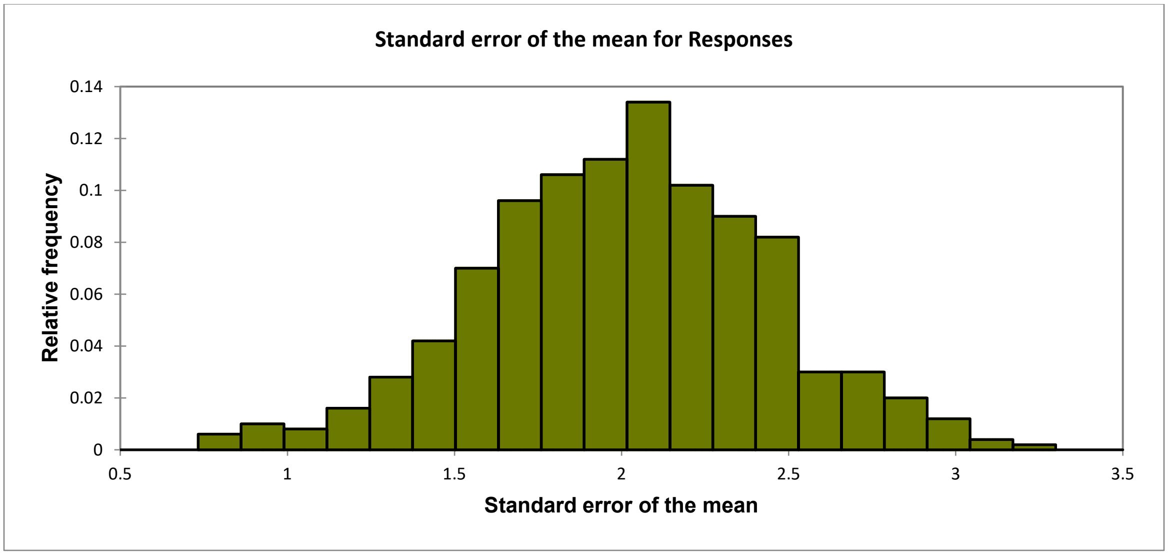
H_0: $p = 0.5$
H_1: $p < 0.5$

Using 500 bootstrap samples for the quantitative variable Responses:

SE = 2.008

95% C.I. = 8.4596 ± 3.9357

Based on the confidence interval of the bootstrap sample, we cannot reject the null hypothesis.

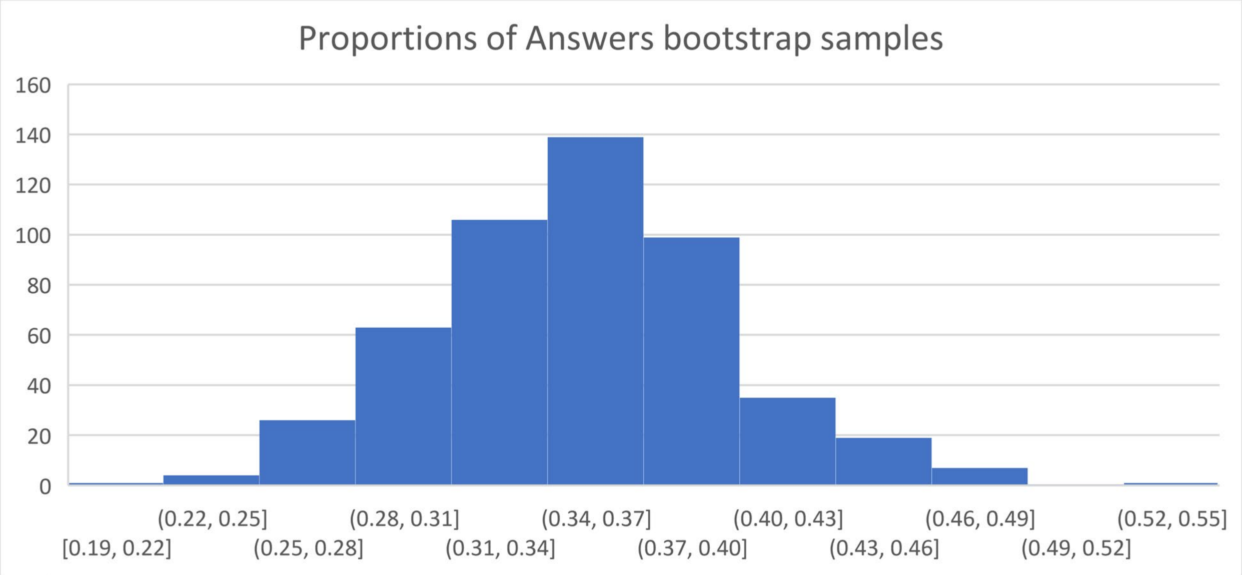


For the bootstrap sample of the qualitative variable Answer using p-hat of none to the other choices:

SE = 0.021

95% C.I. = 0.33 ± 0. 0.042

The z score is -6.768, and the p score from a table is 0, which is significant, so we can reject the null hypothesis that social media hasn't had any effect of shopping habits.



For the categorical variable Answers:

The proportion of none to the other choices (p-hat) = 0.353

SEM = $\sqrt{\frac{(.353)(1-.353)}{7529}}$ = 0.0055

Z-score = $\frac{(.353 - .5)}{.0055}$ = -26.031

P-score from table = 0

Based on the p score being < 50, we can reject the null hypothesis, which is that social media hasn’t had any effect of shopping habits.

For the quantitative variable Responses:

$\bar{x} = 8.4596$

$\mu_0 = 10$

$n = 890$

$s = 60.1088$

$SE = \frac{60.1088}{\sqrt{890}} = 2.0149$

95% CI = 8.4596 ± 3.9492 or [4.5104, 12.4088]

$t = \frac{8.4596 - 10}{2.0149} = -0.7645$

$p = .4448$

Based on both the 95% confidence interval and the p score not being significant at $\alpha = .05$, we cannot reject the null hypothesis of $\mu = 10$. The bootstrap was very similar, the SE had a .0069 difference, and the confidence interval only had a .0135 difference.

The probability that the survey was conducted at a university, with the answer category being twitter.
P(University|Twitter) = $\frac{174}{178}$ = .9775 or 98%
To solve, I took the intersection of University and Twitter, and divided it by the total for Twitter. There is a 98% chance that the answer category was twitter, and the survey was given at a university.

The probability that the survey's answer category was none, given that it was conducted via mobile.
P(None|Mobile) = $\frac{1}{5}$ = .2 or 20%
To solve, I took the intersection of none and mobile, and divided it by the total for mobile. There is a 20% probability that the survey was given on mobile and that the answer category is none.