2019 Poverty Report

Group 2 Team Members:

Yashaswini Thokala,

Aly Kanefield,

Nehemie Joseph

Introduction

Poverty, the state of living without sufficient income to provide for one's needs (or one's family's needs), is a globally recognized issue (United Nations, n.d.). In the United States, several social assistance programs exist to address the US's War on Poverty, first declared in 1964 by the Johnson Administration. Since 2009, the U.S. Census Bureau has annually published The American Community Survey (ACS) Supplemental Poverty Measure (SMP), which provides a detailed snapshot of poverty with consideration of government programs designed to respond to poverty (Census Bureau, October 2021). Our goal is to analyze the 2019 ACS SMP's report on poverty and the respected supplemental poverty programs (Census Bureau, October 2021).

The goal will be achieved by framing the project to answer the following questions:

1. What is the magnitude of aid that supplemental poverty programs provide?
2. Are federal poverty relief programs statistically effective at increasing resources for impoverished people?

Commented [KA1]: We need to follow APA guidelines for headings/elsewhere https://owl.purdue.edu/owl/research_and_citation/apa_style/apa_formatting_and_style_guide/apa_headings_and_seriation.html

Commented [JN2R1]:

Commented [JN3R1]:

Significance

Federal programs, including school lunch subsidies, housing subsidies, and supplemental nutritional assistance, exist to help individuals struggling with notably low income. Understanding the statistical impact of these programs is essential to justifying their cost and supporting their existence. By diving into the 2019 ACS SMP data and estimating the quantitative relationship between household resources and federal programs, this report aims to provide a benchmark for the modern effectiveness of these programs. This information will allow nonprofit organizations to identify holes in the governmental social net where aid is needed. Additionally, this type of analysis can help the US government analyze if programs are effective, if they justify their cost, who they are effective for, and if they need updates to their qualification or payment policies.

Data: Source, Content, and Potential Inaccuracies

This data comes from the 2019 United States Census: American Community Survey Supplemental Poverty Measures (SPM) report. It is considered as an observational study in the scientific world because the population (residents in the USA) that is being studied are not being influenced by researchers (the government) to generate a response. The survey focuses on several facts that occur within the same time period making it a time-series dataset.

The 2019 American Community Survey: Supplemental Poverty Measures is a dataset containing a collection of 46 variables and over 3 billion observations. Some relevant variables that are measured are age, spm_numadults (SPM: Number of Adults), and spm_childcareexpns (SPM: Child Care Expenses). Age and spm_numadults are classified as a quantitative discrete variable meaning the variable depict as only certain values with no intermediate values. spm_childcareexpns is a continuous variable meaning the variable has intermediate values. Each

variable has over 3 billion records which make up the observations. The variables that identify the element (person) which the observations are associated: tax_id (SPM:Tax ID Number), serialno (serial number), spm_fedtax (SPM: Federal Tax Number) and SPM_id. It must be noted that each identification is important in identifying a participant. Two elements can have the same serial no but a different Tax ID which identifies the observation as two separate records.

Descriptive statistics for the content of the sample used in our analysis follows in the methods section.

Errors:

The ACS is a sample survey of the population, that consists of sampling and non-sampling errors. However, due to the efforts to reduce sampling errors that affect the accuracy of the survey any errors within the report are collectively referred to as non-sampling errors by the U.S Census Bureau. ACS's estimates are adjusted to minimize "non-sample" errors by using the Current Population Survey Annual Social report, Economic Supplements (CPS ASEC) report, the Census, and other reports.

Common types of non-sampling errors that occur when conducting the survey:

Over/Under coverage error: Over-coverage is when the selected sample has multiple chances of being selected when they should not have. For example, receiving the two separated id survey twice hence being inputted twice as two different participants when it refers to the same one. Under-coverage is when a participant does not get the chance of being selected when they should have been.

Response error: This occurs when the data is recorded or reported incorrectly either by proxy or by the participant. This error cannot be controlled or truly minimized by ACMS.

Processing error: This occurs when coding the data incorrectly or when there is missing information, or the information of the questionnaire is associated with another question. Experts analyze the edits and content to minimize processing errors.

Voluntary response error: Not completing the survey. Legally, completing the ACS is required by law therefore there is no voluntary response error within the data set. However, due to the size of the survey the law is not heavily enforced therefore some participants do not respond.

Analysis and Modeling Methods

The software programs used to analyze, interpret, and model the data are RStudio, Python, and Excel. Various approaches to statistical modeling are applied for the analysis of data values (i.e., mean, median, standard deviation, quartiles, etc.). Data visualization is employed to show relationships between the variables in the dataset through the creation of bar graphs, histograms, scatter plots, and the like. Finally, regression analysis is done to estimate the direction and magnitude of numerical relationships between variables of interest, namely resources available in a household and factors impacting it, including various expenses and federal poverty programs.

The analysis is done on two subsets of the larger dataset of approximately three million observations. First, for more accurate analysis and better breadth of representation, a stratified random sample was taken of one individual from each household. Because many of the quantitative variables (resources, tax, poverty aid, etc.) describe the household level, not the individual, this sample type was necessary to avoid double counting households and to represent the broadest impact across the population. There are several unique identifiers defining household units, including a household sequence number and a tax unit id. spm_id groups individuals by household units according to the supplemental poverty measures' records. spm_id was chosen as, aside from age, all the variables used in our analysis are recorded on this basis. The stratified random sample consists of one randomly chosen individual from each group (household), resulting in a sample of approximately 85,000 observations. The second subset follows the same process, but only selects from households that fall under the poverty line, leading to a sample of approximately 10,900 observations.

To create the first stratified random sample, the python code groups the whole dataset by spm_id and draws one observation from each unique numerical string representing a household. The second sample is formed using a similar process but iterates over a data frame that has been set to only include households that are officially impoverished. The code run to produce these samples is in Appendix A under the section "Stratified Random Sampling and Other Data Manipulation".

For both samples, the variables of specific interest are Age, Poverty Programs, Total Tax (State and Federal, including FICA and social security), Work/Kid Expenses, and Medical Expenses. Analysis, including descriptive statistics, histograms, and measures of center for these variables follow below.

Descriptive Statistics for the Whole Sample

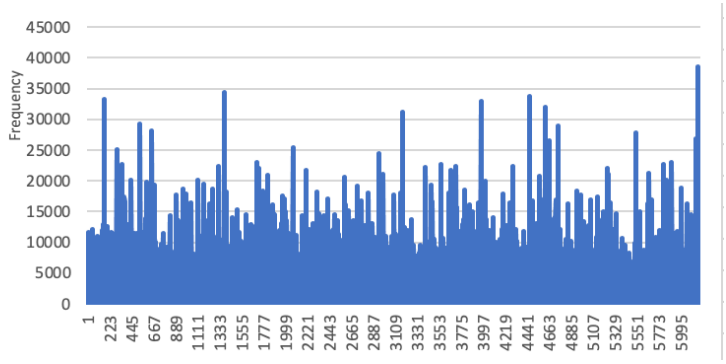
Age:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	30.00	53.00	48.51	66.00	96.00

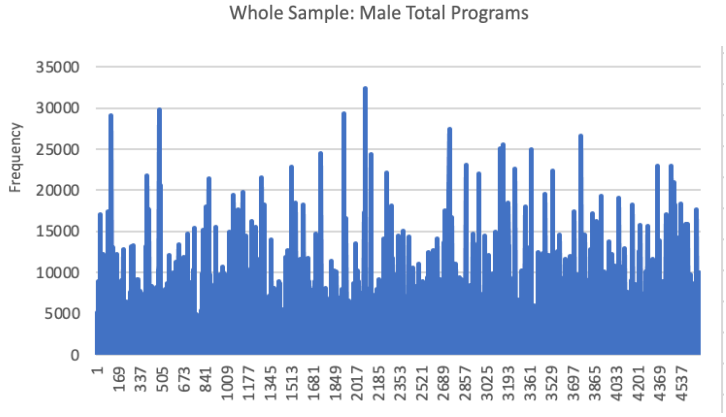
The median age for the whole dataset is 53 years old.

Poverty Programs:

Whole Sample: Female- Total Programs

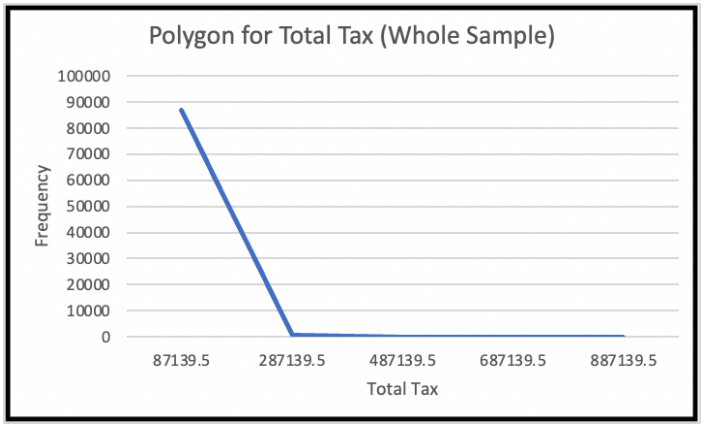


The average amount of total combined resources within the whole sample is $519.23.



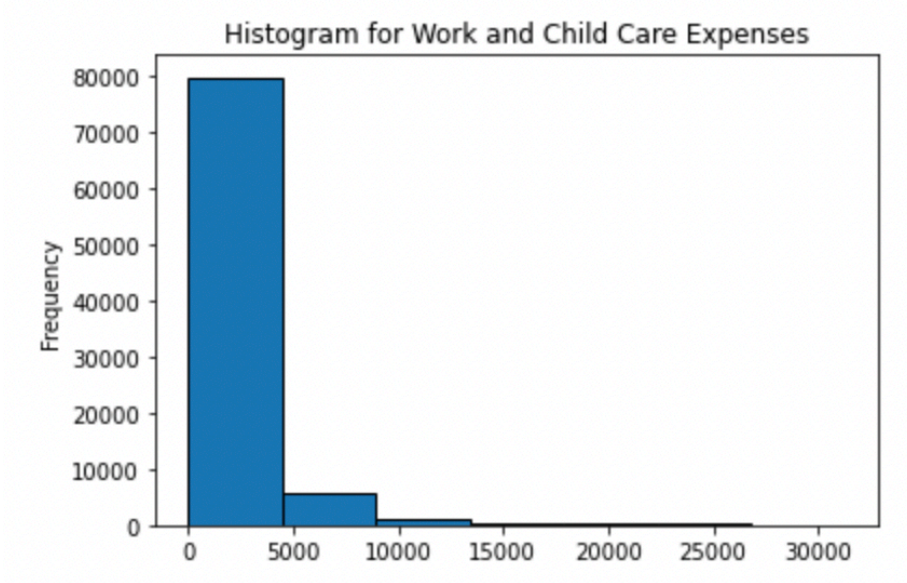
The average amount of total combined resources within the whole sample is $423.74.

Total Tax (State and Federal):



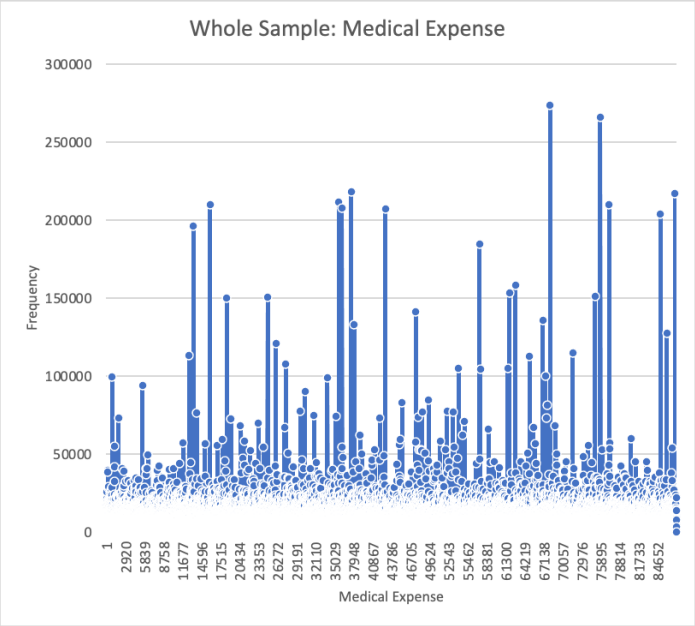
The total tax for the whole sample averages around $87139.5.

Work/Kid Expenses:



The typical range of work expenses is between $0-$5,000.

Medical Expenses:



The average amount of medical expense within the whole sample is $4741.72

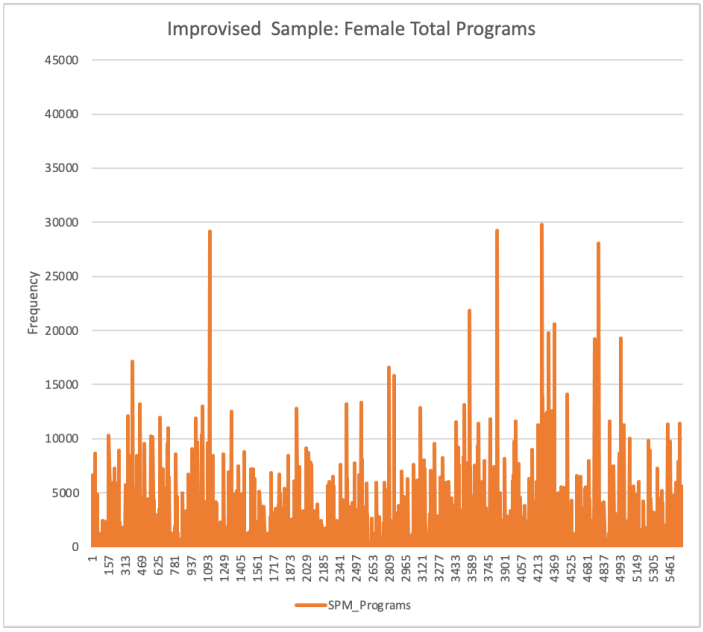
Descriptive Statistics for the Poverty Subsample

Age:

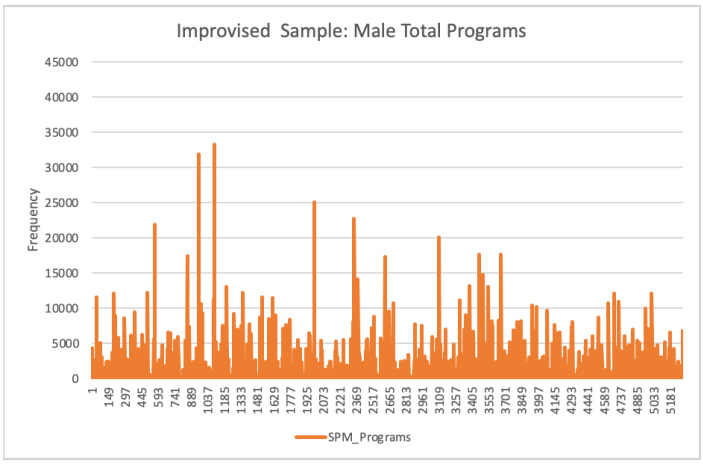
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	22.00	47.00	44.62	64.00	96.00

The median age for the whole dataset is 47 years old.

Poverty Programs:

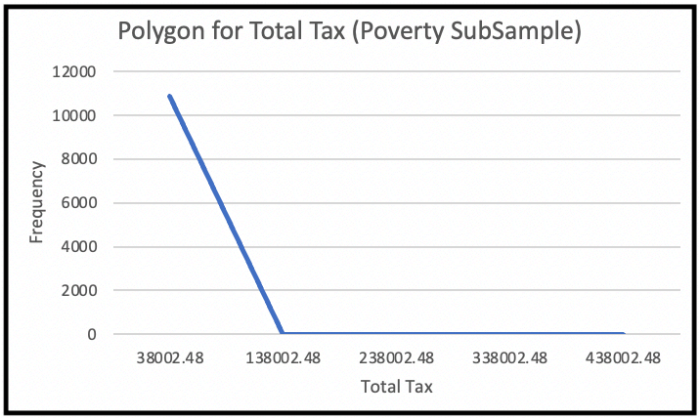


The average amount of total combined resources within the whole sample is $1993.05.



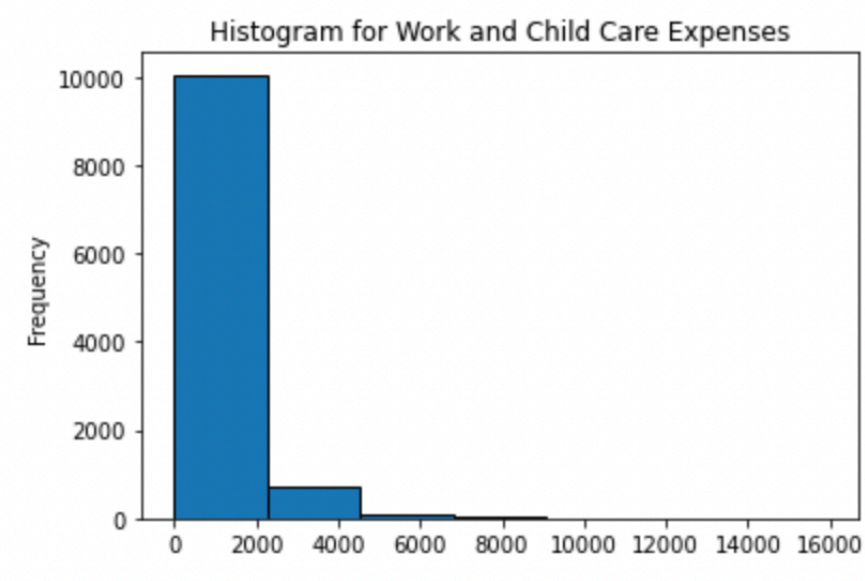
The average amount of total combined resources within the whole sample is $1793.25.

Total Tax (State and Federal):



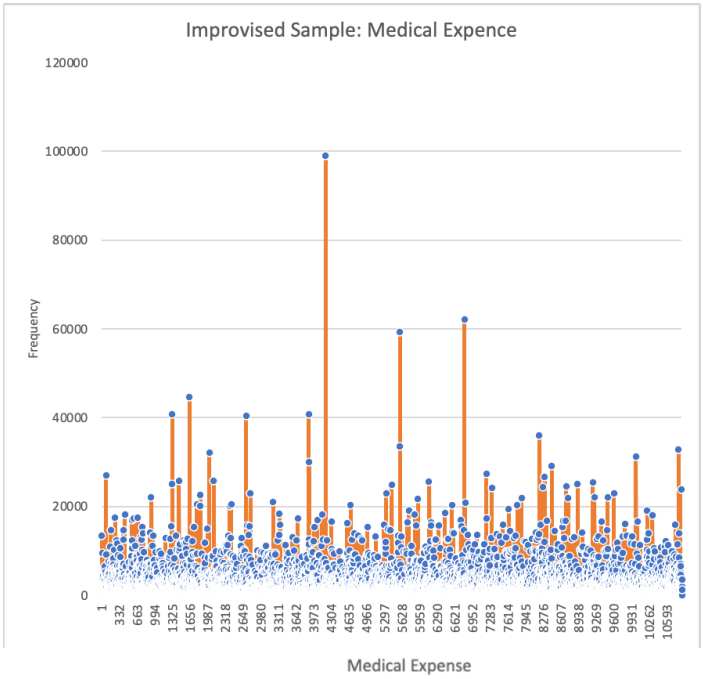
The total tax for the poverty subsample averages around $38002.5.

Work/Kid Expenses:



The typical range of work expenses is between $0-$2,000.

Medical Expenses:



The average amount of medical expense within the whole sample is $1767.61.

Methods Yet to Be Addressed

All code and surplus figures for this section are included in Appendix A and Appendix B, respectively.

Week 2, Chapter 1: Introduction

Page 23. Subset snapshot with replacement.

	spm_fedtax	spm_sttax	spm_schlunch	spm_fedtaxbc	spm_numkids	spm_wcohabit	mar	spm_poor	spm_wui_lt15	spm_eitc
1	38587.50	0.00	0.000	38587.5	0	0	1	0	0	0.00
2	38587.50	0.00	0.000	38587.5	0	0	1	0	0	0.00
3	38587.50	0.00	0.000	38587.5	0	0	5	0	0	0.00
4	340.00	682.34	0.000	340.0	0	0	1	0	0	0.00
5	340.00	682.34	0.000	340.0	0	0	1	0	0	0.00

Page 23. Subset snapshot without replacement.

	spm_schlunch	moop_other	medicare_partb	age	spm_numper	serialno	hispanic	agi	spm_engval	spm_medxpns
1	0.000	6108.70741	2275.2	68	3	1	0	211600.0	0	17535.7360
2	0.000	3050.00000	2275.2	70	3	1	0	211600.0	0	17535.7360
3	0.000	509.05895	0.0	29	3	1	0	58000.0	0	17535.7360
4	0.000	1797.99621	1626.0	80	3	2	0	30400.0	0	6068.1141
5	0.000	763.58843	1626.0	73	3	2	0	30400.0	0	6068.1141

Page 26. Subset snapshot that is a non- random sample.

	sporder	st	puma	wt	age	sex	mar
1	1	12	1208623	171	68	2	1
2	2	12	1208623	156	70	1	1
3	3	12	1208623	185	29	2	5
4	1	18	1800300	68	80	1	1
5	2	18	1800300	67	73	2	1

Page 36. Subset snapshot that using the simple random technique.

	spm_fedtaxbc	race	spm_caphousesub	spm_fica	spm_numper	spm_premium	spm_poor	spm_childcarexpns	spm_resources	st
1	38587.5	1	0.000	18524	3	3317.5696	0	0.000	191633.012	12
2	38587.5	1	0.000	18524	3	3317.5696	0	0.000	191633.012	12
3	38587.5	1	0.000	18524	3	3317.5696	0	0.000	191633.012	12
4	340.0	1	0.000	0	3	152.7177	0	0.000	47509.546	18
5	340.0	1	0.000	0	3	152.7177	0	0.000	47509.546	18

Page 38. Subset snapshot that using the systematic random technique.

	filedate	serialno	sporder	st	puma	wt	age	sex	mar	education	race	hispanic	offpoor	moop_other
28344	20211015	12928	1	39	3900500	75	85	1	1	1	1	0	0	509.05895
59227	20211015	27100	1	25	2503303	71	30	2	5	4	1	0	0	1950.00000
90110	20211015	41285	1	37	3701100	200	76	1	1	2	2	0	1	208.58726
120993	20211015	55369	3	26	2601801	125	20	1	5	0	1	0	1	490.00000

Page 42. Subset snapshot that using the cluster random technique. The unique id that was used is tax_id,

	filedate	serialno	sporder	st	puma	wt	age	sex	mar	education	race	hispanic	offpoor	moop_other
204461	20211015	93436	2	17	1703530	65	64	1	1	3	1	0	0	2747.0942
745625	20211015	340760	2	35	3500100	25	62	2	1	3	4	0	0	1407.9640
1913318	20211015	875309	1	12	1208617	48	60	1	3	4	1	1	0	1730.8004
2108733	20211015	964507	1	37	3704200	65	65	2	5	2	2	0	1	156.4404

Week 3, Chapter 2: Organizing and Graphing Data

Page 10. A frequency distribution is a table which depicts the frequency of the different occurrences in a population or a sample. The following are frequency tables of the qualitative variables of the poverty dataset.

education	freq
College degree	25051
High school degree	18886
Less than a high school degree	6717
Some college degree	20861
Under age 25/NIU	16034

Fig.: Frequency Distribution table for Education

spm_wui_lt15	freq
Has UI under 15	413
No UI under 15	87136

Fig.: Frequency Distribution Table for SPM unit of number of unrelated individuals under 15 years old

spm_poor	freq
In Poverty	13038
Not in Poverty	74511

Fig.: Frequency Distribution Table for SPM Poverty Status

Page 11. Relative frequency depicts the popularity of each outcome of the category with the entire population. The following tables depict the relative frequency and percentage of the above-mentioned qualitative variables.

Education	freq	Rel. Freq.	percent
College degree	25051	0.28613691	28.6
High school degree	18886	0.21571920	21.6
Less than a high school degree	6717	0.07672275	7.7
Some college degree	20861	0.23827799	23.8
Under age 25/NIU	16034	0.18314315	18.3

Fig.: Relative Frequency and Percentage for Education

Unrelated Under 15	freq	Rel. Freq.	percent
Has UI under 15	413	0.004717358	0.5
No UI under 15	87136	0.995282642	99.5

Fig.: Relative Frequency and Percentage for SPM unit for number of unrelated individuals under 15 years old

SPM Poor Status	freq	Rel. Freq.	percent
In Poverty	13038	0.1489223	14.9
Not in Poverty	74511	0.8510777	85.1

Fig.: Relative Frequency and Percentage for SPM Poverty Status

Page 15. Bar graphs are visual displays of categorical variables which consist of rectangular bars whose heights are determined by data proportional to the variables they depict. Pareto charts contain both bars and a line graph where each individual value is represented by the bars while the cumulative

values are represented by the line graph. The following are the bar graphs and pareto charts for qualitative variables.

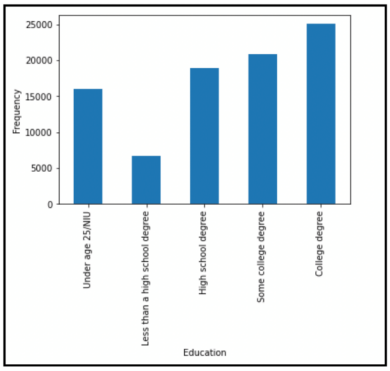


Fig.: Bar Graph for Education

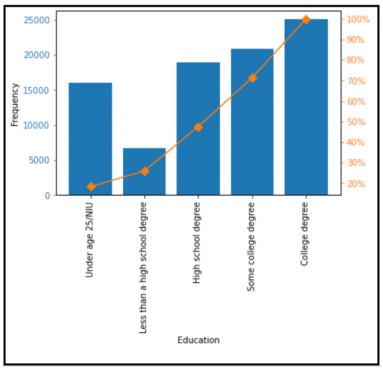


Fig.: Pareto Chart for Education

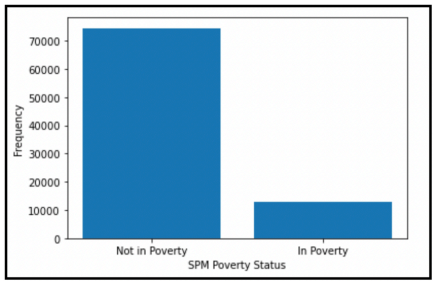


Fig.: Bar Graph for SPM Poverty Status

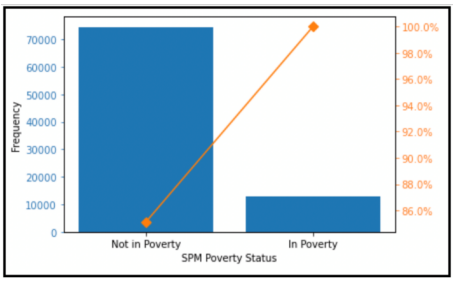


Fig.: Pareto Chart for SPM Poverty Status

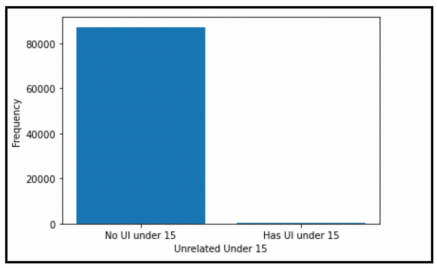


Fig.: Bar Graph for Unrelated Under 15

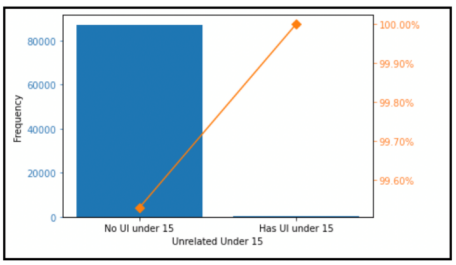


Fig.: Pareto Chart for Unrelated Under 15

Page 19-23. The following are frequency distribution tables for the quantitative variables of the dataset.

age	freq
0 to less than 20	11292
20 to less than 40	20065
40 to less than 60	22678
60 to less than 80	27520
80 to less than 100	5994

Fig.: Frequency Distribution Table for Age

agi	freq
-4800 to less than 495200	86874
495200 to less than 995200	650
995200 to less than 1495200	24
1495200 to less than 1995200	1

Fig.: Frequency Distribution Table for Adjusted Gross Income

wkccxpns	freq
0 to less than 10000	85873
10000 to less than 20000	1309
20000 to less than 30000	353
30000 to less than 40000	14

Fig.: Frequency Distribution Table for SPM's unit capped work and childcare expenses

medxpns	freq
0 to less than 50000	87461
50000 to less than 100000	58
100000 to less than 150000	13
150000 to less than 200000	7
200000 to less than 250000	8
250000 to less than 300000	2

Fig.: Frequency Distribution Table for SPM unit's Medical Out-of-Pocket (MOOP) and Medicare Part B subsidy

Page 29. A histogram is like a bar graph except that in a histogram the range of values are divided into a series of equal intervals and then find the frequency of values falling under each of these

intervals. Finally, bars or bins are constructed adjacent to each other to represent the frequency of values falling in each of these intervals. The following are examples of histograms for quantitative variables in the dataset.

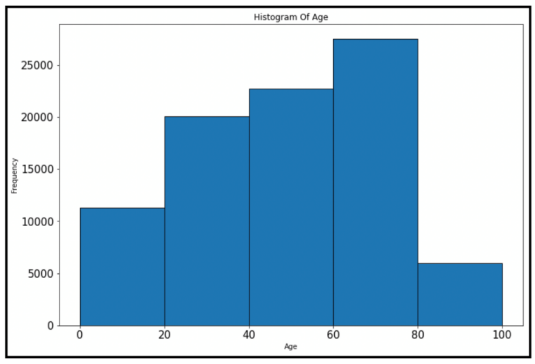


Fig.: Histogram for Age

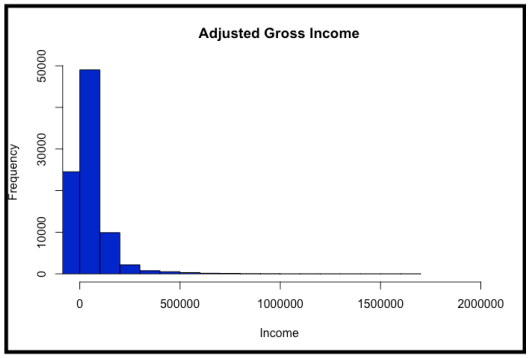


Fig.: Histogram for Adjusted Gross Income

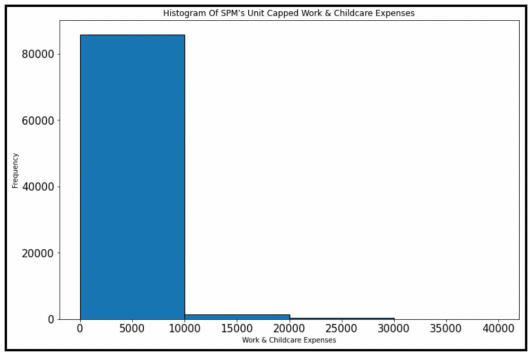


Fig.: Histogram for SPM's unit Capped Work & Childcare Expenses

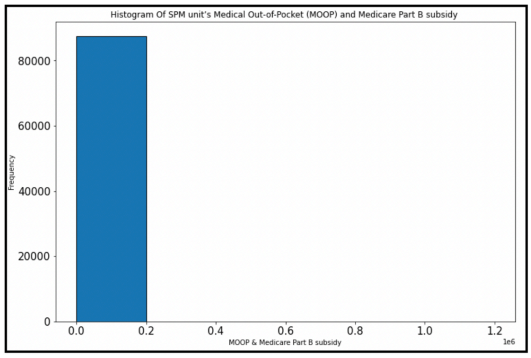


Fig.: Histogram for SPM unit's Medical Out-of-Pocket (MOOP) and Medicare Part B subsidy

Page 30. A polygon graph is a graph that is drawn by joining the midpoints of each interval created for the values. The height of these midpoints in the graph represents the frequency. The following are the polygon graphs for quantitative variables in the dataset.

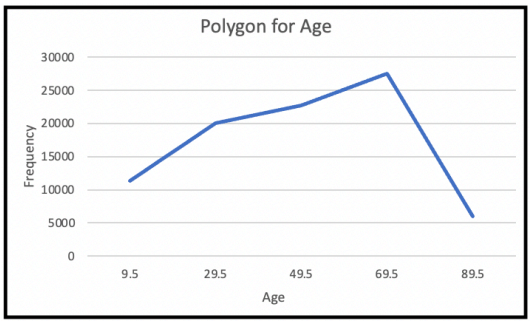


Fig.: Polygon Graph for Age

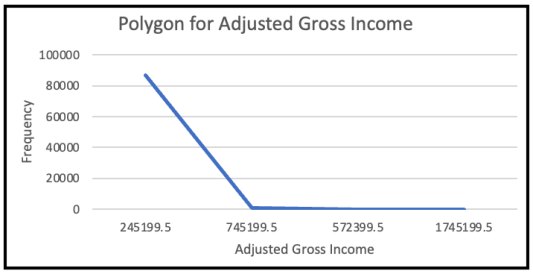


Fig.: Polygon Graph for Adjusted Gross Income

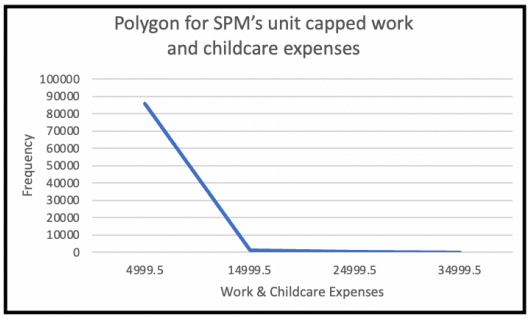


Fig.: Polygon Graph for SPM's unit Capped Work & Childcare Expenses

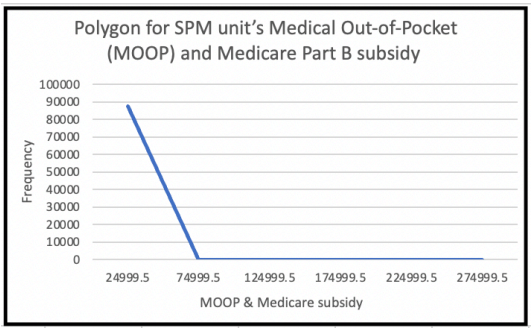


Fig.: Polygon Graph for SPM unit's Medical Out-of-Pocket (MOOP) and Medicare Part B subsidy

Page 37 – 39. Cumulative Frequency is the total sum of the frequency of one class and the frequencies of all classes below it. Relative frequency is the ratio of the frequency of each class to the total size. Cumulative Relative Frequency is the sum of the relative frequency of one class with the relative frequency of all classes below it. Cumulative Percentage is the cumulative frequencies of each class divided by 100.

Age	Frequency	Cum. Freq	Cum. Rel. Freq	Cum. Percentage
0 to less than 20	11292	11292	0.1290	12.8979
20 to less than 40	20065	31357	0.3582	35.8165
40 to less than 60	22678	54035	0.6172	61.7197
60 to less than 80	27520	81555	0.9315	93.1535
80 to less than 100	5994	87549	1.0000	100

Table: Cumulative Frequency Distribution Table for Age

Agi	Frequency	Cum. Freq	Cum. Rel. Freq	Cum. Percentage
-4800 to less than 495200	86874	86874	0.992290032	99.22900319
495200 to less than 995200	650	87524	0.999714446	99.97144456
995200 to less than 1495200	24	87548	0.999988578	99.99885778
1495200 to less than 1995200	1	87549	1	100

Table: Cumulative Frequency Distribution Table for Adjusted Gross Income (agi)

Work & Childcare Expenses	Freq	Cum. Freq.	Cum. Rel. Freq	Cum. Percentage
0 to less than 10000	85873	85873	0.980856435	98.08564347

10000 to less than 20000	1309	87182	0.995808062	99.58080618
20000 to less than 30000	353	87535	0.99984009	99.98400895
30000 to less than 40000	14	87549	1	100

Table: Cumulative Frequency Distribution Table for SPM's unit Capped Work & Childcare Expenses

MOOP & Medicare Part B subsidy	Freq	Cum. Freq.	Cum. Rel. Freq	Cum. Percentage
0 to less than 50000	87461	87461	0.998994849	99.89948486
50000 to less than 100000	58	87519	0.999657335	99.96573347
100000 to less than 150000	13	87532	0.999805823	99.9805823
150000 to less than 200000	7	87539	0.999885778	99.98857782
200000 to less than 250000	8	87547	0.999977156	99.99771556
250000 to less than 300000	2	87549	1	100

Table: Cumulative Frequency Distribution Table for SPM unit's Medical Out-of-Pocket (MOOP) and Medicare Part B subsidy

Page 41 - 44. The histogram for the quantitative variable 'Age' is right-skewed and hence is said to be positive.

The histogram for the quantitative variable 'Adjusted Gross Income' is normally distributed.

The histogram for the quantitative variable 'SPM's unit Capped Work & Childcare Expenses' is right-skewed and hence is said to be positive.

The histogram for the quantitative variable 'SPM unit's Medical Out-of-Pocket (MOOP) and Medicare Part B subsidy' is right-skewed and hence is said to be positive.

Page 50-52. A stem and leaf display is a form of table used to display data. The 'stem' is on the left side of a vertical line and consists of the first digits of a series of values. The 'left' is written on the right side and shows the remaining of the number.

The following are the ages of 10 randomly selected people from the Poverty Census dataset.

76 59 56 86 33 70 37 55 29 60

Stem-and-leaf display:

```
1   |
    |
2   | 9
    |
3   |  3  7
    |
4   |
    |
5   |  5  6  9
    |
6   |  0
    |
7   |  0  6  9
    |
8   |  6
    |
9   |
```

Page 54 – 57. A dot plot is a representation of data points plotted on a simple graph, with the number of dots for each value representing the frequency of occurrence of the value in the data.

The following are the ages of 10 randomly selected people from the Poverty Census dataset.

69 34 41 69 53 50 69 34 29 64

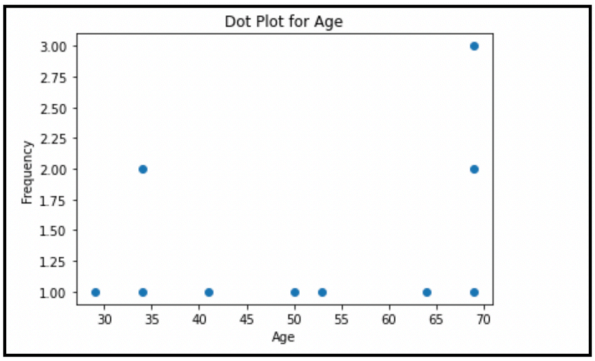


Fig.: Dot Plot for 10 randomly selected ages

Week 4, Chapter 3: Numerical Descriptive Measures

For all analysis below, the stratified random sample is used. Because many values are listed for the whole household and included for each individual, analysis of the whole dataset includes many repeats.

Page 7. The means for the relevant variables are as follows: spm_resources = 65199.17, age = 48.51, SPM_Programs = 473.51, SPM_Tax = 18221.53, spm_capwkccxpns = 2484.12, spm_medxpns = 4741.71.

Page 8-13. The 10% trimmed mean of total resources is $65,361.72 and the mathematical average is $77,149.14.

Page 9. The median of the relevant variables are spm_resources = 48075.83, age = 53.00, SPM_Programs = 0.00, SPM_Tax = 7921.60, spm_capwkccxpns = 2065.47, spm_medxpns = 3455.62.

Page 10. Mean and median are both measures of center. The difference between them comes down to whether they are resistant to the presence of outliers. The mean is biased; the median is not. The mean is calculated by the sum of all variables divided by the number of variables (n). The sum is biased by the magnitude of the outliers while n remains the same regardless of the value of the individual numbers. Thus the mean is biased by the values of the outliers. The median is unbiased. It is merely the value of the number situated at index n/2 of a

ranked dataset or the average of the 2 center numbers of an even dataset. Since outliers fall at the far ends of a ranked dataset, the figures used in the median are not impacted by them.

Page 11. The mode for the variables included in this analysis are as follows:

Resources	Age	PovProg	TotTax	WrkKidXpns	MedXpns
0	61	0	0	2065.47	0

Resources, TotTax, and MedXpns have a mode of zero likely because the individual values are distinctly unique to each household. If rounded to the nearest 1,000 or 10,000 dollars, there would likely be 1-3 modes representing common income levels. The value for work and kid expenses represents the capped value that SPM records this level at. PovProg is 0 because more households do not receive benefits than those that do at different levels. The mode for Age, 61, is likely because this age represents individuals with the time and interest to respond to the ACS SPM survey.

Page 12. All the relevant variables are unimodal. The only variables in the whole sample that are multimodal are the unique identifiers, which represents the households with the greatest number of individuals.

Page 13. The weighted mean is calculated for the first ten rows of the original dataset, which includes 10 people in three households according to their tax id number. The average adjusted gross income per person for these ten people is $84,200.00.

Page 19-21. The two randomly chosen variables are spm_eitc and spm_schlunch, which represent the households' earned income tax credit and their school lunch subsidy, respectively. The mean indicates that on average, the yearly earned income tax credit for each household in this sample is $249.83, and the school lunch subsidy received is $100.96. The median and mode for both is 0 for the uniquely distinct characteristic mentioned above.

	Spm_eitc	Spm_schlunch
Mean	249.83	100.96
Mode	0	0
Median	0	0

Page 23. The ranges for the relevant variables are spm_resources = 1049751.32, age = 96.00, SPM_Programs = 38572.93, SPM_Tax = 806490.85, spm_capwkccxpns = 31276.31, spm_medxpns = 273899.31.

Page 24. Variance and standard deviation describe how near to the mean the values of a dataset are. The variance measures the average squared distance that all of the values are away from the mean. The standard deviation is the square root of the variance, i.e. the average distance the values are away from the mean. The variance and standard deviation values are as follows:

Variance:		Standard Deviation:
spm_resources	4392962509.83	66279.43
age	522.96	22.87

SPM_Programs	3067828.69	1751.52
SPM_Tax	1202310662.65	34674.35
spm_capwkccxpns	7646773.65	2765.28
spm_medxpns	34238213.77	5851.34

Page 28-31. The correlation coefficient for the number of kids is 202.94%, and the correlation coefficient allotted school lunch subsidy is 351.33%. Standard deviations are quite large, even relative to their respective means.

Page 47. The average age of this sample is 48.50 and the standard deviation is 22.87. Find the minimum percentage of individuals that fall within 1.5 standard deviations of the mean, which is between the ages of 10.19 and 78.81 years.

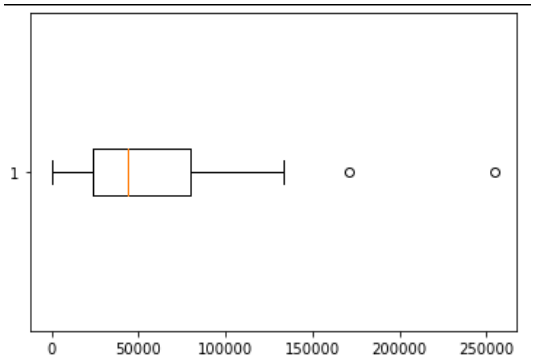
Chebyschev's theorem states that at least 56% of the data will fall within 1.5 standard deviations of the mean, so 56% of this sample are within 10.19 and 78.81 years of age.

Page 49-50. The empirical rule only applies to distributions that are normal or approximately normal. None of the distributions of the variables in this sample resemble a normal distribution, so this exercise is not applicable here. Histograms of the whole sample are included in Figure 1 of Appendix B.

Page 52-53. 20 Random percentiles and percentile ranks for adjusted gross income are in Figure 2 of Appendix B.

Page 54. The quartiles and IQR for the adjusted gross income of 20 observations are included in Figure 3 of Appendix B.

Page 56-61. Below is a box and whisker plot of total household resources for 30 random observations of this sample.



Week 6, Chapter 4: Probability

Page 9. An experiment would be to choose one random member of the population and check whether this person is officially in poverty by looking at which number is in the 'offpoor'

column. The outcomes are 1 – in poverty or 0 – not in poverty, and the sample space is {1 – in poverty, 0 – not in poverty}.

Page 10. If the experiment is to choose two people and see if either of them are in poverty, an example of an event would be that the outcome is one is in poverty and one is not {1,0}. Because this is just one outcome for the experiment, this example is also a simple event.

Page 12. An example of a compound event would be that at least one of the two individuals is in poverty. That compound event includes the outcomes {1,0}, {0,1}, and {1,1}.

Page 25. An example of a mutually exclusive event is the event that a person is both under the age of 5 and has a college degree (education = 5). Independent events are those where the probability of one event does not affect the other, or P(A|B) = P(A) or P(B|A) = P(B). In this dataset, the event being from AZ and being under 18 are independent. P(age<18) = 635,717/3,088,232 = 21% and the P (age<18|from AZ) is 14,266/67,649 = 21%. Dependent events are those where the above condition is not met. For example, the probability that an individual is in poverty (offpoor = 1) and that they have less than a high school degree (education = 1) is not independent. P(offpoor==1) = 323,369/3,088,232 = 10.47% and the P(offpoor=1|education=1) = 46,825/218,767 = 21.40%.

Page 26. If event A is the outcome of being officially in poverty (offpoor = 1), then the compliment of event A (A bar) is the outcome of not being officially in poverty. (offpoor = 0).

Page 29. If event A is having an adjusted gross income above the sample first quartile of $7,325, and event B is having an agi less than the sample third quartile of $60,000, then the intersection of event A and B is every individual with an income level between $7,325 and $60,000. Event A includes agi levels above 60k while B includes levels below 7.3k, and A union B only includes those in between.

Page 33. If event A is that the individual's sex is female and event B is that the sex is male, the union of these events contains all individuals who are either male or female, which, in accordance with the classifications available in this data, includes the whole data set.

Week 8, Chapter 7: Sampling Distributions

Page 24.

1.

Q. The mean age for all 3088232 people in the poverty dataset is 42 years and the standard deviation is 23.75703 years. Let $\bar{x}$ be the mean age for a random sample of people selected from the dataset. Find the mean and standard deviation of x for a sample size of

(i) 10,000 (ii) 500,000 (iii) 1,000,000

Ans.:

The mean of the age is 42 years.

The standard deviation of the age is 23.75703 years.

The mean is the same regardless of the sample size, hence mean for each sample size is 42.

For 10000,

$$\frac{n}{N} = \frac{10000}{3,088,232} = 0.0032 \leq 0.05$$

Hence, the standard deviation for a sample of 10,000 is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{24}{\sqrt{10000}} = \mathbf{0.24}$$

For 500000,

$$\frac{n}{N} = \frac{500000}{3,088,232} = 0.16 > 0.05$$

Hence, the standard deviation for a sample of 500,000 is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{24}{\sqrt{500000}} \sqrt{\frac{3088232-500000}{3088232-1}}$$

$$=0.0339 \sqrt{\frac{2588232}{3088231}} = \mathbf{0.03103}$$

For 1,000,000 -

$$\frac{n}{N} = \frac{1000000}{3,088,232} = 0.323 > 0.05$$

Hence, the standard deviation for a sample of 1,000,000 is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{24}{\sqrt{1000000}} \sqrt{\frac{3088232-1000000}{3088232-1}}$$

$$=0.024 \sqrt{\frac{2088232}{3088231}} = \mathbf{0.0197}$$

2.

2. Assume that the number of persons in a household taken for the poverty census are approximately normally distributed with a mean of 3 and a standard deviation of 1.70979. Find the probability that the mean number of people, x, of a random sample of 500 households will be between 2.8 and 2.9.

The shape of the sampling distribution of $\bar{x}$ is approximately normal because the population is approximately normally distributed.

$\mu_{\bar{x}} = \mu = 3$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1.70979}{\sqrt{500}} = 0.07646$

T compute the probability that the value of x calculated for one randomly drawn sample of 500 is between 2.8 and 2.9; that is,

$$P(2.8 < x < 2.9)$$

This probability is given by the area under the normal distribution curve for x between the points x = 4 and x = 10. The first step in finding this area is to convert the two x values to their respective z values.

The z values for x = 2.8 and x = 2.9 –

For x = 2.8: $z = \frac{2.8-3}{0.07646} = -2.615$

For x = 2.9: $z = \frac{2.9-3}{0.07646} = -1.307$

The probability that x is between 2.8 and 2.9 is given by the area under the standard normal curve between z = −2.615 and z = −1.307, which is obtained by subtracting the area to the left of z = −2.615 from the area to the left of z = −1.307. Thus, the required probability is

$P(2.8 < x < 2.9) = P(-2.615 < z < -1.307)$

$= P(z < -1.307) - P(z < -2.615)$

$=$ **0.091**

Therefore, the probability that the mean number of persons of a sample of 500 households will be between 2.8 and 2.9 is **0.091**.

3.

7 8 9 10 11 12 13 14 16

4956 2425 1116 523 229 60 39 16 20

According to the data set the number of households that include at least seven kids has a percentage of

$$\frac{no.of\ rows\ with\ atleast\ 7\ kids}{total\ no\ of\ rows} = \frac{9384}{3088232} = \quad 0.0031$$

Assuming that this is true for all the households that includes at least seven kids. Let $\hat{p}$ be the proportion of households in a random sample of 10000 which include at least seven kids. Find the mean and standard deviation of $\hat{p}$ and describe the shape of its sampling distribution.

Let p be the proportion of all households with at least seven kids,

p = 0.0031 q = 1 – p = 1 – 0.0031 = 0.9969 and n = 10000

The mean of the sampling distribution of $\hat{p}$ is

$$\mu_{\hat{p}} = p = 0.0031$$

The standard deviation of $\hat{p}$ is

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.0031)(0.9969)}{10000}} = 0.00055$$

The value of np and nq are

np = 10000(0.0031) = 31

nq = 10000(0.9969) = 9969

Since both the values, np and nq, are greater than 5, the central limit theorem can be applied and hence the sampling distribution of $\hat{p}$ is approximately normal with a mean of **0.0031** and standard deviation of **0.00055**.

4.

P (Households having 4 to 9 adults) = $\frac{2505949}{3088232}$ = 0.8114

81.14% of households have 4 to 9 adults. Suppose this is true for all current households. Let $\hat{p}$ be the proportion in a random sample of 10000 households. Find the probability that 80.5% to 81% of households in this sample have 4 to 9 adults.

n = 10000, p = 0.8114, and q = 1 − p = 1 − 0.8114 = 0.1886, where p is the proportion of households have 4 to 9 adults

The mean of the sample proportion $\hat{p}$ is $\mu_{\hat{p}} = p = 0.8114$

The standard deviation of $\hat{p}$ is $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.8114)(0.1886)}{10000}} = 0.00391$

As np = 10000(0.8114) = 8114 and nq = (10000) (0.1886) = 1886 are greater than 5, central limit theorem could be applied to say that the data is approximately normally distributed.

p (0.805 < x < 0.81)

$z = \frac{\bar{p}-p}{\sigma} = \frac{0.805-0.8114}{0.00391} = -1.636$

$z = \frac{\bar{p}-p}{\sigma} = \frac{0.81-0.8114}{0.00391} = -0.358$

p (-1.636 < z < -0.358) = 0.3602 – 0.0509

= **0.3093**

Week 9, Chapter 8: Estimations of mean and proportions

This is the 90 % confidence interval of the whole dataset focusing on agi.

```
60867.80 61048.85
```

Inferential Analysis: Ordinary Least Squares Regression

OLS (ordinary least squares) linear regression is applied to deepen the understanding of the relationship between variables herein. OLS regressions describe the relationship between one or more independent variables (x axis) and a dependent variable (y axis) by minimizing the sum of the squared error terms of the data to a line of best fit. OLS regressions provide estimates of the size and direction of impact that an independent variable (regressor) has on the dependent variable.

The regression in this report is derived from the goal of this report: to understand 1) the magnitude of aid that supplemental programs provide, and 2) if federal poverty relief programs are statistically effective at increasing resources for impoverished people. The equation used is as follows:

Total Resources=α + x Age + x PovPrograms + x TotTax + x WrkKidXpns + x MedXpns + u

It is run on the data from both random samples outlined earlier in this section. The results are discussed below.

Results

The output from the regressions run on the two samples are as follows:

Total Resources(Whole Sample) = 23,872.06 + 79.92 Age + 0.25 PovPrograms - 1.73 TotTax - 1.87 WrkKidXpns - 0.23 MedXpns + u

R^2 = 86.99% Adjusted R^2 = 86.99%

Total Resources(in Poverty) = 5066.88 + 10.00 Age + 1.36 PovPrograms - 1.37 TotTax - 4.79 WrkKidXpns + 0.71 MedXpns + u

R^2 = 64.85% Adjusted R^2 = 63.83%

All the coefficients are statistically significant above a 0.1% level, except that of Age which is statistically significant at the 5% level. The high level of significance on the coefficients indicates that none are highly correlated at an individual level. If collinearity was present, the high level of correlation would likely make one of the two variables insignificant. The coefficient on one would capture the effects of the other, rendering the latter insignificant. While there may be some degree of multicollinearity, the model still functions well. Its effectiveness is evidenced by the f-statistic for joint significance, which has a p-value of $2.2x10^{-16}$ for both models. A highly statistically significant f-statistic for all the regressors shows that the likelihood of the coefficients of all the regressors simultaneously equaling zero, as they would for a model with little to no explanatory power, is highly unlikely. This has nearly a 0% probability in this case.

Additionally, both models have relatively high adjusted R^2 values. Adjusted R^2 measures the percentage of variation around the mean in total resources that is explained by the change in the independent variables: age, poverty programs, taxes, and work, child, and medical expenses. Adjusted R^2 is different from that of R^2 in that it adjusts for the increase in degrees of freedom that comes with the addition of more regressors. The adjusted R^2 value for the whole sample is higher, at 86.99%, than that of the subset of that sample including only those below the poverty line, an R^2 of 63.83%. This is because there is a higher level of correlation between the independent variable and the regressors in the whole sample than there is between them in the poverty only sample. Conceptually speaking, this speaks to the complicated nature of poverty. More variables are needed to accurately describe the relationship between the level of resources available and what impacts it. Future considerations to improve both levels of adjusted R^2 include the integration of qualitative variables, such as education, race, ethnicity, location, or gender.

Conclusion

The goal stated in the introduction of this report was to answer the following questions:

1. What is the magnitude of aid that supplemental programs provide?
2. Are federal poverty relief programs statistically effective at increasing resources for impoverished people?

The coefficient on the PovPrograms variable answers both. For the whole sample, a $1 increase in poverty program aid is estimated to increase resources available to the household by $0.25. Naturally, the impact of these programs for the poverty-only sample is much higher, with an estimated increase in household resources of $1.36 for every $1 increase in federal poverty programs. These answer the first question. For both samples, the coefficient is highly statistically significantly different from 0 at a confidence level higher than 99%. This answers the second question: yes, federal poverty relief programs do increase resources available to those living under the poverty line the United States.

These models are highly simplified representations of complicated individual scenarios. Future work would benefit from the inclusion of more data on an individual level. Additionally, this data is only from 2019. Time series analysis, including integration of ARMA or GARCH models to ensure qualifying assumptions, could reveal more about the effectiveness of these programs and the impact of other variables over time, and speak more directly about their ability to improve outcomes for impoverished people overall. Regardless, the evidence points to the programs' ability to effectively provide aid, thus justifying their existence.

Appendix A

This appendix contains the code used to manipulate and analyze the data across Python, R Studio, and Excel.

Week 2

Sampling- Python- #library(survey):

- *#Sampling with Replacement*

  sam_replacemnt <- sample(ACS_SPM, size= 10, replace = TRUE, prob = NULL)

- *#Sampling without Replacement*

  sam_woreplacemnt<- *sample(ACS_SPM, size= 10, replace = FALSE, prob = NULL)*

- *#Random Sampling*

  sample_replacemntrandom<- sample(ACS_SPM, size = 10)

- *#Non-random Sampling*

  sample_replacemntnonrandom<-head(ACS_SPM)

- *#Simple Random Sampling Technique*

  set.seed (45)

  sam_replacemntsimple<- sample(ACS_SPM, size = 10, replace = FALSE, prob = NULL)

- *#Systematic Random Sampling Technique*

  #define function to obtain systematic sample obtain_sys = function(N,n){ k = ceiling(N/n) r = sample(1:k, 1) seq(r, r + k*(n-1), k)} #obtain systematic sample sys_sample_df = df[obtain_sys(nrow(df), 100),] #view first six rows of data frame head(sys_sample_df)

Source: https://www.statology.org/systematic-sampling-r/

- *#Sampling Cluster Sampling Technique*

clusters <- sample(unique(ACS_SPM$tax_id), size=4, replace=T)

cluster_sample <- ACS_SPM[ACS_SPM$tax_id %in% clusters,]

Source: https://www.statology.org/cluster-sampling-r/

Week 3

Frequency tables of Qualitative Variables

- *#Frequency Distribution Table for Education (education) (RStudio)*

  count(censusdatadf, 'education')

- *#Frequency Distribution Table for SPM unit has unrelated individual (UI) under 15 years old (spm_wui_lt15) (RStudio)*

  count(censusdatadf, 'spm_wui_lt15')

- *#Frequency Distribution Table for SPM Poverty Status (spm_poor) (RStudio)*

  count(censusdatadf, 'spm_poor')

Relative Frequency and Percentage Table of Qualitative Variables

- *#RELFREQ EDUCATION (RStudio)*

```
relfreqed <- table(censusdatadf$education)/length(censusdatadf$education)
edrelf <- cbind.data.frame(relfreqed)
colnames(edrelf) <- c('Education', 'Rel. Freq.')
df <- count(censusdatadf, 'education')
edrelf$freq <- df$freq
library(epiDisplay)
per <- tab1(censusdatadf$education, sort.group = "ascending", cum.percent = FALSE)
edrelf$percent <- per$Percent
per <- cbind.data.frame(per)
per <- head(per, -1)
edrelf$percent <- per$output.table.Percent
edrelf <- edrelf[, c("Education", "freq", "Rel. Freq.", "percent")]
edrelf
```

- *#RELFREQ Unrelated Individual under 15 years old (RStudio)*

```
relfreqlt15 <- table(censusdatadf$spm_wui_lt15)/length(censusdatadf$spm_wui_lt15)
lt15relf <- cbind.data.frame(relfreqlt15)
colnames(lt15relf) <- c('Unrelated Under 15', 'Rel. Freq.')
df <- count(censusdatadf, 'spm_wui_lt15')
lt15relf$freq <- df$freq
library(epiDisplay)
per <- tab1(censusdatadf$spm_wui_lt15, sort.group = "ascending", cum.percent = FALSE)
lt15relf$percent <- per$Percent
per <- cbind.data.frame(per)
```

```
per <- head(per, -1)
lt15relf$percent <- per$output.table.Percent
lt15relf <- lt15relf[, c("Unrelated Under 15", "freq", "Rel. Freq.", "percent")]
lt15relf
```

- *#RELFREQ SPM POOR (RStudio)*

```
relfreqspmp <- table(censusdatadf$spm_poor)/length(censusdatadf$spm_poor)
spmprelf <- cbind.data.frame(relfreqspmp)
colnames(spmprelf) <- c('SPM Poor Status', 'Rel. Freq.')
df <- count(censusdatadf, 'spm_poor')
spmprelf$freq <- df$freq
library(epiDisplay)
per <- tab1(censusdatadf$spm_poor, sort.group = "ascending", cum.percent = FALSE)
spmprelf$percent <- per$Percent
per <- cbind.data.frame(per)
per <- head(per, -1)
spmprelf$percent <- per$output.table.Percent
spmprelf <- spmprelf[, c("SPM Poor Status", "freq", "Rel. Freq.", "percent")]
spmprelf
```

Bar Graphs and Pareto Charts for Qualitative Variables

- *Bar Graph for Education (Python)*

```
import matplotlib.pyplot as plt
df.plot.bar(x='Education', y='Frequency', legend = False)
plt.xticks(rotation=90)
plt.ylabel('Frequency')
```

- *Bar Graph for Unrelated Individuals under 15 (Python)*

```
import matplotlib.pyplot as plt
```

```
plt.bar(df['Unrelated Under 15'], df['Frequency'])
plt.xlabel('Unrelated Under 15')
plt.ylabel('Frequency')
```

- *Bar Graph for SPM Poverty Status (Python)*

```
plt.bar(df['SPM Poverty Status'], df['Frequency'])
plt.xlabel('SPM Poverty Status')
plt.ylabel('Frequency')
```

- *Pareto Chart for Education (Python)*

```
df["cumpercentage"] = df["Frequency"].cumsum()/df["Frequency"].sum()*100
fig, ax = plt.subplots()
ax.bar(df['Education'], df["Frequency"], color="C0")
ax2 = ax.twinx()
ax2.plot(df['Education'], df["cumpercentage"], color="C1", marker="D", ms=7)
ax2.yaxis.set_major_formatter(PercentFormatter())
plt.xticks(rotation = 50, horizontalalignment="right")
ax.tick_params(axis="y", colors="C0")
ax2.tick_params(axis="y", colors="C1")
plt.show()
```

- *Pareto Chart for Unrelated Individuals under 15 (Python)*

```
df["cumpercentage"] = df["Frequency"].cumsum()/df["Frequency"].sum()*100
from matplotlib.ticker import PercentFormatter
fig, ax = plt.subplots()
ax.bar(df['Unrelated Under 15'], df["Frequency"], color="C0")
ax2 = ax.twinx()
ax2.plot(df['Unrelated Under 15'], df["cumpercentage"], color="C1", marker="D", ms=7)
ax2.yaxis.set_major_formatter(PercentFormatter())
ax.set_xlabel('Unrelated Under 15')
```

```
ax.set_ylabel("Frequency")
ax.tick_params(axis="y", colors="C0")
ax2.tick_params(axis="y", colors="C1")
plt.figure(figsize=(40000,10000))
plt.show()
```

- *Pareto Chart for SPM Poverty Status (RStudio)*

```
from matplotlib.ticker import PercentFormatter
df["cumpercentage"] = df["Frequency"].cumsum()/df["Frequency"].sum()*100
fig, ax = plt.subplots()
ax.bar(df['SPM Poverty Status'], df["Frequency"], color="C0")
ax2 = ax.twinx()
ax2.plot(df['SPM Poverty Status'], df["cumpercentage"], color="C1", marker="D", ms=7)
ax2.yaxis.set_major_formatter(PercentFormatter())
ax.set_xlabel('SPM Poverty Status')
ax.set_ylabel("Frequency")
ax.tick_params(axis="y", colors="C0")
ax2.tick_params(axis="y", colors="C1")
plt.figure(figsize=(40000,10000))
ax.set_xlabel('SPM Poverty Status')
plt.show()
```

Frequency Distribution Tables for Quantitative Variables

- *#FreqDist For Age (RStudio)*

```
age <- c('0 to less than 20', '20 to less than 40', '40 to less than 60', '60 to less than 80', '80 to less than 100')
freq <- c(11292, 20065, 22678, 27520, 5994)
df <- data.frame(age, freq)
df
```

- *#FreqDist for Adjusted Gross Income (RStudio)*

```
agi <- c('-4800 to less than 495200', '495200 to less than 995200', '995200 to less than 1495200', '1495200 to less than 1995200') #, '1995200 to less than 2495200', '2495200 to less than 2995200')

freq <- c(86874, 650, 24, 1)

df<- data.frame(agi, freq)

df
```

- *#FreqDist for Work & Childcare Expenses (RStudio)*

```
wkccxpns <- c('0 to less than 10000', '10000 to less than 20000', '20000 to less than 30000', '30000 to less than 40000')

freq <- c(85873, 1309, 353, 14)

df <- data.frame(wkccxpns, freq)

df
```

- *#FreqDist for MOOP & Medicare Part B subsidy (RStudio)*

```
medxpns <- c('0 to less than 50000', '50000 to less than 100000', '100000 to less than 150000', '150000 to less than 200000', '200000 to less than 250000', '250000 to less than 300000')

freq <- c(87461, 58, 13, 7, 8, 2)

df <- data.frame(wkccxpns, freq)

df
```

Histogram for Quantitative Variables

- *Histogram for Age (Python)*

```
Povertydf['age'].plot(kind='hist', bins=5, title='Histogram Of Age', rot=0,
    grid=False, edgecolor = "black", figsize=(12,8), fontsize=15, range = [0, 100])

    plt.xlabel("Age");

    plt.ylabel("Frequency");
```

- *Histogram for Adjusted Gross Income (RStudio)*

```
hist(censusdatadf$agi, xlim=c(-4800,1995200), main="Adjusted Gross Income",
    xlab="Income", col = c('medium blue'))
```

- *Histogram for Work and Child Care Expenses (Python)*

```
Povertydf['spm_capwkccxpns'].plot(kind='hist', bins=4, title='Histogram Of SPM's Unit Capped Work & Childcare Expenses', rot=0, grid=False, edgecolor = "black", figsize=(12,8), fontsize=15, range=[0, 40000])

plt.xlabel("Work & Childcare Expenses");

plt.ylabel("Frequency");
```

- *Histogram for MOOP & Medicare Part B subsidy (Python)*

```
Povertydf['spm_medxpns'].plot(kind='hist', bins=6, title='Histogram Of SPM unit's Medical Out-of-Pocket (MOOP) and Medicare Part B subsidy ', rot=0, grid=False, edgecolor = "black", figsize=(12,8), fontsize=15, range = [0, 1200000])

plt.xlabel("MOOP & Medicare Part B subsidy", fontsize = 10);

plt.ylabel("Frequency");
```

Dot Plot for Quantitative Variable 'Age'

```
age = {29: [0, 0, 1], 34: [0, 1, 2], 41: [0, 0, 1], 50: [0, 0, 1], 53: [0, 0, 1], 64: [0, 0, 1], 69: [1, 2, 3]}

age1 = pd.DataFrame(age, columns = [29, 34, 41, 50, 53, 64, 69])

df2=age1.melt()

mask = df2.value > 0.5

plt.scatter(df2.variable[mask], df2.value[mask])

plt.xlabel("Age")

plt.ylabel("Frequency")

plt.title("Dot Plot for Age")
```

Stratified Random Sampling and Other Data Manipulation

NOTE: This section, Week 4, and Week 6 were written in the same file under one main() function. If copying the code back into a .py file, copy this section, Week 4, and Week 6 to complete the main() function if intending to run the whole file at once or wishing to make it callable. Additionally, the bolded titles for Week 4 & Week 6 should be deleted before attempting to run the code. Some comments can be returned to code to produce desired output.

```
# -*- coding: utf-8 -*-
```

```
"""
Created on Fri Nov 19 16:34:34 2021

@author: alyka
"""

import pandas as pd
import matplotlib.pyplot as plt
from scipy import stats

#Checking where this file is saved so I know where to save csv
#from os.path import abspath
#current_file_name = __file__
#full_path = abspath(current_file_name)
#print(full_path)

def main():

    pd.set_option('float_format', '{:f}'. format)
        #set to read output as whole values before decimals instead of 4.30e^10 or whatever
    #use nrows to load first 250,000 if processing time is too slow
    full_df=pd.read_csv('Census2019Poverty.csv')
    onefromspm=pd.read_csv('onefromspm.csv')
    #df.isnull().values.any()
        #check for any null values - NONE

#create column for sum of total taxes and sum of poverty programs to understand combined
significance
```

```
    full_df['SPM_Programs'] = full_df['spm_snapsub'] +
full_df['spm_caphousesub']+full_df['spm_schlunch']+full_df['spm_engval']+full_df['spm_wicval
']

    full_df['SPM_Tax'] = full_df['spm_fica'] + full_df['spm_fedtax'] + full_df['spm_sttax']


    onefromspm['SPM_Programs'] = onefromspm['spm_snapsub'] +
onefromspm['spm_caphousesub']+onefromspm['spm_schlunch']+onefromspm['spm_engval']+on
efromspm['spm_wicval']

    onefromspm['SPM_Tax'] = onefromspm['spm_fica'] + onefromspm['spm_fedtax'] +
onefromspm['spm_sttax']

    #Save onefromspm2.csv to have these columns

    onefromspm.to_csv('onefromspm2.csv')


#Stratified random sample of one person from each tax id

    one_each = full_df.groupby('tax_id', group_keys=False).apply(lambda x: x.sample(1))

    one_each.to_csv('oneHH2.csv')

#Stratified random sample of one person from each spm id

    onefromspm = full_df.groupby('spm_id', group_keys=False).apply(lambda x: x.sample(1))

    onefromspm.to_csv('onefromspm2.csv')

#Stratified random sample of one person from each spm id but everyone is impoverished

    pov_df = full_df.loc[full_df['offpoor']==1]

        #Creates a df that has only impoverished individuals

    povonespm = pov_df.groupby('spm_id', group_keys=False).apply(lambda x: x.sample(1))

    povonespm.to_csv('povspmfixed.csv')

#creating a df for relevant variables

    rel_df = onefromspm.filter(items = ['spm_resources', 'age', 'SPM_Programs', 'SPM_Tax',
'spm_capwkccxpns', 'spm_medxpns'])


#Adjusting var for number of individuals in the household unit

    onefromspm['adjRes'] = onefromspm['spm_resources']/onefromspm['spm_numper']
```

```
    onefromspm['adjProg'] = onefromspm['SPM_Programs']/onefromspm['spm_numper']
    onefromspm['adjTax'] = onefromspm['SPM_Tax']/onefromspm['spm_numper']
    onefromspm['adjMed'] = onefromspm['spm_medxpns']/onefromspm['spm_numper']
    onefromspm['adjwrkkid'] = onefromspm['spm_capwkccxpns']/onefromspm['spm_numper']
    onefromspm.to_csv('onefromspmAdj.csv')
    #Some lines for verification
    #Read first row/first 5 rows to ensure it's working
    #r1=df.iloc[0]
    #print(r1)
    #print(df.head())
```

Week 4 (Python) Ch 3 Numerical Descriptive Measures

```
#CHAPTER 3/WEEK 4 NUMERICAL DESCRIPTIVE MEASURES
#7 get average of each variable
    mean_all = full_df.mean()
    print("MEANS BELOW")
    rel_mean = rel_df.mean()
    print(rel_mean)
    #Double checked one var with R to verify - is good


#8 and 13 try the k% trimmed mean
    resource_trimmedavg = stats.trim_mean(full_df.spm_resources, .1)
    resource_mathavg = full_df['spm_resources'].mean()
    #The 10% trimmed mean is $65,361.72 and the mathematical average is $77,149.14


#9 get median of each var
    med_all = full_df.median()
    print("MEDIANS BELOW")
    rel_med = rel_df.median()
```

```
    #print(med_all)

#11 get mode for each var
    mode_all = full_df.mode('index')
    print("MODE BELOW")
    relevant_modes = rel_df.mode('index')
    #print(mode_all)
    #NaN rows exist to fill space in unimodal columns
    #modes fill rows for columns with more than one mode

#13 weighted mean
#doing average agi per person according to tax unit
#can't use group by bc i won't actually be using the tax_id in the calcs
#but rather a count per tax_id
    df_ten = pd.read_csv('Census2019Poverty.csv', nrows = 10)
    count_per_unit= df_ten.pivot_table(columns=['tax_id'], aggfunc='size')
        #counts duplicates of unique figures in tax_id column
    df_count = count_per_unit.reset_index()
        #converts count pivot table to df
    df_count.rename(columns = {0: 'ppl_count'}, inplace = True)

    #Various methods to find agi according to tax id
    #df_ten.query('tax_id==10000000')['agi']
        #Returns all instances
    #df.loc[df_ten['tax_id'] == 10000000, 'agi'].iloc[0]
        #returns one float64
    #best iterable version in loop below
```

```
    unit_agi = []
        #Create empty list to fill with agi to later add to df_count
    for index, row in df_count.iterrows():
        current_tax = df_count['tax_id'][index]
            #finds the tax id number in df with duplicate counts
        index_agi = df_ten['agi'][full_df['tax_id'] == current_tax].iloc[0]
            #Searches df with agi for agi value
        unit_agi.append(index_agi)
            #adds agi to agi list
    df_count['id_agi'] = unit_agi
        #adds unit_agi column to df_count


    #Weighted mean adjusted gross income per person
    df_count['agixcount'] = df_count['id_agi'] * df_count['ppl_count']
    weighted_mean_agi_per_person = df_count['agixcount'].sum()/df_count['ppl_count'].sum()
    print("Weighted mean of adjusted gross income per person in tax unit for first ten people
BELOW")
    print(weighted_mean_agi_per_person)


#19-21 mean, median, and mode of two variables
    df_two_ran_var = full_df.sample(2, axis=1)
        #Creates a df of two randomly selected variables
    two_mean = df_two_ran_var.mean()
    two_med = df_two_ran_var.median()
    two_mode = df_two_ran_var.mode()
    print("Two var means are:")
    print(two_mean)
    print('Two var medians are:')
    print(two_med)
```

```
    print("The two var modes are:")
    print(two_mode)

#23 range for all variables
    df_max = full_df.max()
    df_min = full_df.min()
    df_range = pd.DataFrame({'max': df_max, 'min': df_min})
    df_range['range'] = df_range['max'] - df_range['min']

    relmax = rel_df.max()
    relmin = rel_df.min()
    rel_range = pd.DataFrame({'max': relmax, 'min': relmin})
    rel_range['range'] = rel_range['max'] - rel_range['min']


#24 variance and standard deviation for all variables
    df_stdev = full_df.std()
    df_var = full_df.var()
    rel_stdev = rel_df.std()
    rel_var = rel_df.var()
    print(rel_var)
    print(rel_stdev)

#28-31 coefficient of variation for number of kids and school lunch subsidy
    stratsd = onefromspm.std()
    stratmean = onefromspm.mean()
    coeff_var_kids = (stratsd['spm_numkids']/stratmean['spm_numkids'])*100
    coeff_var_schlunch = (stratsd['spm_schlunch']/stratmean['spm_schlunch'])*100
```

```
    #astype percent? OR just make sure to clarify when you print
    print("The coefficient of variation for kids and lunch are as follows: ")
    print(coeff_var_kids)
    print(coeff_var_schlunch)
    print("coefficient of variation represented as percentages")

#47 Chebyshev - any distribution
    age_mean = onefromspm['age'].mean()
    print("The mean age for this data is " + str(age_mean) + ".")
    age_std = onefromspm['age'].std()
    print("The standard deviation is " + str(age_std) + ".")
    #find the minimum percentage of individuals in this sample with an agi between [mu +- 2std]

#49-50 Empirical rule - bell curve - age ignoring under 18 roughly bell shaped
#But not close enough to be valid- All histograms will print below to demonstrate none are normally distributed enough
    #plt.hist(full_df['age'], bins= 20)
        #Just age histogram
    onefromspm.hist(bins=20, grid = False, xlabelsize=0, ylabelsize=0, figsize = (9,9))
        #histograms of all columns



#53 calculate percentile and percentile rank of 20 observations using agi
    full_df['agi % rank'] = (full_df.agi.rank(pct=True) * 100)
    full_df['agi percentile'] =  full_df['agi % rank'].round()
    ran20 = full_df.sample(n=20)
    print("Below are the percentiles of 20 random members from the data")
    print(ran20['agi percentile'])
```

```
    print("Below are the percentile ranks (in percentages) of the 20 random members from the
data")
    print(ran20['agi % rank'])


    onefromspm['agi % rank'] = (onefromspm.agi.rank(pct=True) * 100)
    onefromspm['agi percentile'] =  onefromspm['agi % rank'].round()
    ran20 = onefromspm.sample(n=20)
    print("Below are the percentiles of 20 random members from the data")
    print(ran20['agi percentile'])
    print("Below are the percentile ranks (in percentages) of the 20 random members from the
data")
    print (ran20['agi % rank'])


#54 quartiles and IQR of 20 observations of agi
    Q1 = ran20['agi'].quantile(q=.25)
    Q2 = ran20['agi'].quantile(q=.5)
    Q3 = ran20['agi'].quantile(q=.75)
    Q4 = ran20['agi'].quantile(q=1)
    IQR = Q3 - Q1
    print(IQR)
    LowFence = Q1 - (IQR*1.5)
    TopFence = Q3 + (IQR*1.5)



#56-61 box and whisker of 30 observations using one variable
    boxplt = plt.boxplot(onefromspm['spm_resources'].sample(n=30), vert=False,
showfliers=True)
        #will generate random 30 sample each time
```

Week 6 (Python)

NOTE: This section is intended to be copied into the same .py file as above to keep the main() function whole.

```
#CHAPTER 6


#Independent
    under18 = full_df.age<18
        #binary series of individuals under 18
    fromAZ = full_df.st==4
        #binary from AZ or not
    indep_df = pd.concat([under18,fromAZ], axis=1)
        #creates df of above series
    indep_df.sum()
        #produces count of total who are under 18 and total who are from az
    indep_df['both'] = count_df.sum(axis=1)
    indep_df.value_counts()
        #gives count of both from az and minor, from az not minor, not from az minor, and neither


#Dependent
    nohs = full_df.education==1
        #binary series of individuals with less than a high school degree
    offinpov = full_df.offpoor==1
        #binary officially in pov or not
    countdep_df = pd.concat([nohs,offinpov], axis=1)
        #creates df of above series
    countdep_df.sum()
        #produces count of total who are under pov line and have less than hs dip
    countdep_df['both'] = countdep_df.sum(axis=1)
    countdep_df.value_counts()
        #Gives count in 4-way table
```

```
if __name__ == '__main__':
    main()
```

Week 9

Regression Analysis (R Studio)

```
setwd("C:/Users/alyka/OneDrive/Documents/Clark/Grad School/Intermediary Statistical Modeling/R Files")
library(MASS)
library(ISLR)


cenpov <- read.csv("Census2019Poverty.csv")
  #full data set
onespm <- read.csv("onefromspm.csv")
  #stratified random sample
onespm2 <- read.csv("onefromspm2.csv")
  #Stratified random sample with columns that sum
  #state, federal, federal insurance into SPM_Tax
  #and various poverty programs into SPM_Programs
povspm <- read.csv("povspm.csv")
povspm2 <- read.csv("povspm2.csv")
  #these are the same as above but have only impoverished individuals
adjspm <- read.csv('onefromspmAdj.csv')
  #divides values at household level by how many individuals are in each household



#The following programs have a negative effect but are measured positively
#Multiply by -1 and overwrite so the coefficients in regressions show correct signs
onespm2$SPM_Tax <- -1 * onespm2$SPM_Tax
onespm2$spm_capwkccxpns <- -1 * onespm2$spm_capwkccxpns
```

```
onespm2$spm_medxpns <- -1 * onespm2$spm_medxpns

povspm2$SPM_Tax <- -1 * povspm2$SPM_Tax
povspm2$spm_capwkccxpns <- -1 * povspm2$spm_capwkccxpns
povspm2$spm_medxpns <- -1 * povspm2$spm_medxpns

adjspm$adjMed <- -1 * adjspm$adjMed
adjspm$adjTax <- -1 * adjspm$adjTax
adjspm$adjwrkkid <- -1 * adjspm$adjwrkkid

#scatter plots to view relationships
#more linear than logistic though not perfectly linear
plot(onespm2$SPM_Tax, onespm2$spm_resources)
points(onespm2$spm_medxpns, onespm2$spm_resources, col="blue")
points(onespm2$spm_capwkccxpns, onespm2$spm_resources, col="red")
  #negative var
plot(onespm2$SPM_Programs, onespm2$spm_resources)
points(onespm2$age, onespm2$spm_resources, col="green")
  #positive var



#Create a data set without outliers to test if model has more explanatory power with them in
Q1 <- quantile(onespm2$spm_totval, .25)
Q3 <- quantile(onespm2$spm_totval, .75)
IQR <- IQR(onespm2$spm_totval)
no_out_spm2 <- subset(onespm2, onespm2$spm_totval> (Q1 - 1.5*IQR) &
onespm2$spm_totval < (Q3 + 1.5*IQR))
```

#Many models were run for trial and error, but the models applied in the regression analysis herein are:

SPM9a = lm(spm_resources~ age + SPM_Programs + SPM_Tax + spm_medxpns + spm_capwkccxpns, data = onespm2)

SPM9a

summary(SPM9a)

#runs on whole stratified random sample

SPM9apov = lm(spm_resources~ age + SPM_Programs + SPM_Tax + spm_medxpns + spm_capwkccxpns, data = povspm2)

SPM9apov

summary(SPM9apov)

#runs on those for whom offpoor=1 in the stratified random sample

SPM9aNoOut = lm(spm_resources~ age + SPM_Programs + SPM_Tax + spm_medxpns + spm_capwkccxpns, data = no_out_spm2)

SPM9NoOut

summary(SPM9aNoOut)

#runs on data with no outliers

#lowers R^squared so chose not to use

SPM9aAdj = lm(adjRes~ age + adjProg + adjTax + adjMed + adjwrkkid, data = adjspm)

SPM9aAdj

summary(SPM9aAdj)

#runs on data that had been manipulated so that variables captured at household level were divided by the

#number of individuals in the household

#not effective

#needs more nuanced approach with application to individual variables

#which should be divided by adults only, which by children only, etc.

Week 9

confidence internval for the mean of AGI

```
sample.mean <- mean(ACS_SPM$agi)
sample.n <- length(ACS_SPM$agi)
sample.sd <- sd(ACS_SPM$agi)
sample.se <- sample.sd/sqrt(sample.n)
alpha = 0.10
degrees.freedom = sample.n - 1
t.score = qt(p=alpha/2, df=degrees.freedom,lower.tail=F)
  margin.error <- t.score * sample.se
    lower.bound <- sample.mean - margin.error
  upper.bound <- sample.mean + margin.error
print(c(lower.bound,upper.bound))
```

Source: https://bookdown.org/logan_kelly/r_practice/p09.html

Appendix B

This appendix contains surplus figures, including graphs and tables, not included in the main body of the report.

Figure 1: Histograms of All Variables for Sample

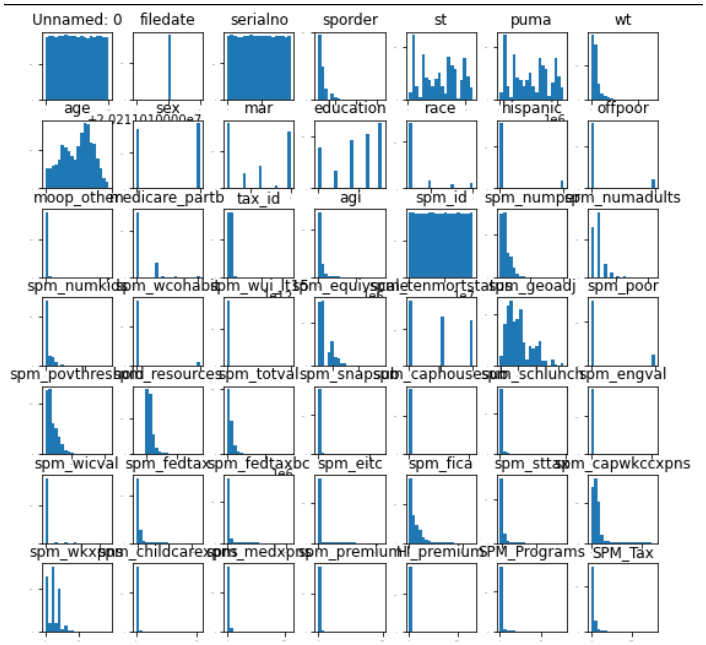


Figure 2: 20 Random Percentiles and Percentile Ranks from Sample

Index	23410	12813	68075	71627	46637	59609	61139	77154	67693	53007
agi % rank	44.49	62.51	46.70	58.20	67.98	14.06	74.98	36.24	14.06	80.94
agi percentile	44	63	47	58	68	14	75	36	14	81
Index	75105	12527	71724	49077	16276	59030	10553	59277	73184	83847
agi % rank	28.80	66.49	28.02	75.35	59.83	63.09	30.29	94.29	38.77	84.14
agi percentile	29	66	28	75	60	63	30	94	39	84

Figure 3: Quartiles and IQR of Adjusted Gross Income for 20 Random Observations

Q1	Q2	Q3	Q4	IQR
$7,325.00	$ 39,810.00	$60,000.00	$182,000.00	$52,675.00

Appendix C

References

Census Bureau, U. S. (2021, October 22). *ACS Supplemental Poverty Measures (SPM) research files: 2009 to 2019*. Census.gov. Retrieved November 1, 2021, from https://www.census.gov/data/datasets/time-series/demo/supplemental-poverty-measure/acs-research-files.html.

Census Bureau, U. S. (2021, September 14). *Income and poverty in the United States: 2020*. Census.gov. Retrieved November 1, 2021, from https://www.census.gov/library/publications/2021/demo/p60-273.html.

Fox, L. et al. (2020, September). *The Supplemental Poverty Measure: 2019*. Retrieved November 1, 2021, from https://www.census.gov/content/dam/Census/library/publications/2020/demo/p60-272.pdf.

Matthews, D. (2021, November 25). *Everything you need to know about the war on poverty*. The Washington Post. Retrieved December 11, 2021, from https://www.washingtonpost.com/news/wonk/wp/2014/01/08/everything-you-need-to-know-about-the-war-on-poverty/.

United Nations. (n.d.). *Goal 1. End Poverty in All Its Forms Everywhere*. United Nations. Retrieved November 1, 2021, from https://www.un.org/sustainabledevelopment/poverty/