

Some General Thoughts about the Discussion Questions:

Mean and Median and Mode

The mean and median are measures of central tendency for quantitative variables. The mean is calculated using all of the observed values of the variables and weights them by their relative frequencies. Thus, it is sensitive to all of the values and is useful when you want a mathematical expectation (prediction) of the values (important later on). The median, on the other hand, is not sensitive to the extreme values (e.g., Bill Gates' income) that may occur and is a good measure of a typical value from a distribution. The median is the 50th percentile, thus half of the measured values fall above it and half below.

For categorical or qualitative variables neither the mean nor the median are appropriate statistics. What would they mean for a qualitative or categorical variable?

The mean and median will differ if a distribution is asymmetric, but will be similar to each other for a symmetric distribution. This holds true regardless of the presence of outliers at the high or low end (as long as they are balanced in symmetric distributions). We will soon discuss normal distributions, one type of symmetric distribution. However be warned that having the same mean and median does not make a distribution normal, this is only one of the necessary conditions.

Standard Deviation, Range, Maximum and Minimum

The standard deviation is one measure of the "spread" of a distribution for a quantitative variable. It is the square root of the mean of the squared differences between the values and their mean (i.e., the square root of the variance). The scale of the standard deviation is the same as that for the mean (whereas the variance is in squared units and its size may be harder to think about.) The standard deviation and variance, like the mean, take into account all of the measured values. Because the difference between each value and the mean is squared, extremely high or low values can have an even greater impact on the variance and standard deviation than on the mean.

IF you know the type of distribution that the values follow (e.g., normal) then you can say something about what fraction of the values fall within one standard deviation, or two standard deviations, etc. **BUT if you don't know the distribution type, then you cannot say exactly what fraction of the values fall within one standard deviation or two standard deviations, etc.**

The range, maximum and minimum values are measures of spread that reflect the extremes of the distribution. It is worth considering whether the highest and lowest value are corrected recorded values that just happen to be very high or low (e.g., Bill Gates' income), or whether they are errors in the dataset that should be corrected if that is possible and deleted if that is not possible. The percentiles of the distribution are another tool to describe the spread of values and valuable as they are not sensitive to the extreme values.

Histograms and Bar Charts

Histograms give a lot of information about the distribution for a numerical variable. You will want to notice whether the distribution is sharply peaked or fairly flat which shows how similar households or individuals are in the values of the variable. For example, how do incomes vary across households in our sample?

For a categorical variable (such as commuting method), a bar chart shows the relative frequency of each category. You will want to notice which categories are common and which are not and whether this information is consistent with other information you have.

The Data Set

The data set you used for this assignment is a telephone survey of Washington State. If our population of interest is households in Washington State then the data are great for answering our questions. However if we want to extrapolate to households in other states we would need to understand how those populations of households were similar or different than those in Washington and how we could account for those differences. Also, there may not be enough households in our sample to get good information about some subpopulations (e.g., people who bike to work or female-headed households in Spokane).

Here is an example of what a couple paragraphs from the 2002 WSPS summarizing some descriptive statistical information might look like for this assignment:

This study uses the 2002 Washington Population Survey data to examine the relationship between household size and income for Washington state households. The data included information on many demographic factors and housing characteristics from a telephone survey of 17,437 individuals from 6,842 households in Washington State.

Figure 1: Household size

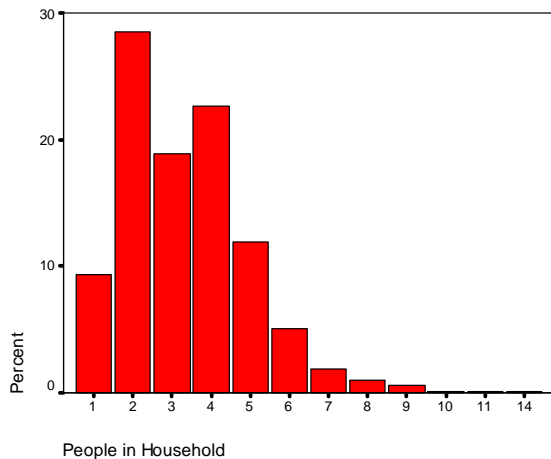


Figure 1 shows the distribution of household size and Figure 2 shows household income. Households in the sample had on average 3.3 members and varied from 1 to 14. Less than 5 percent of households had 6 or more members.

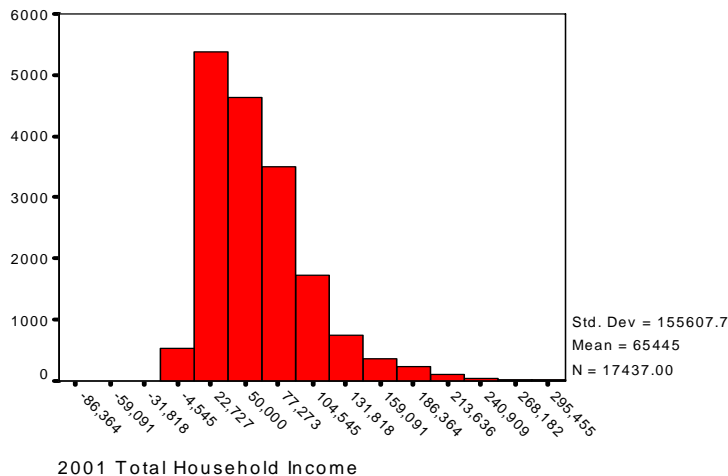


Figure 2: Household Income

Households averaged \$65,444 in income, though income varied greatly. (The standard deviation is very large, \$155,607, and is primarily attributable to some extremely high income responses¹). Most households had total income of \$120,000 or less. [Note: it'd be better to have percentages on Y axis here.]

size and income, perhaps using another measure of income (i.e., income per household member) to allow consideration of equity issues.
PLEASE LOOK AT EXAMPLE FINAL PAPERS ON THE WEB FOR ADDITIONAL EXAMPLES.

In the future research I will look at the relationship between household

¹ Two people reported household incomes of more than \$10 million

People in Household

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 1	1628	9.3	9.3	9.3
2	4968	28.5	28.5	37.8
3	3285	18.8	18.8	56.7
4	3955	22.7	22.7	79.3
5	2085	12.0	12.0	91.3
6	882	5.1	5.1	96.4
7	322	1.8	1.8	98.2
8	168	1.0	1.0	99.2
9	99	.6	.6	99.7
10	20	.1	.1	99.9
11	11	.1	.1	99.9
14	14	.1	.1	100.0
Total	17437	100.0	100.0	

Variable 1 – Number of People in Household (possible descriptive output)

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
People in Household	17437	1	14	3.32	1.627
Valid N (listwise)	17437				

Variable 2 – Total Household Income for 2001 (possible descriptive output)

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
2001 Total Household Income	17437	0	10600000	65444.93	155607.681
Valid N (listwise)	17437				