

PBAF 527

Answer Sheet for Assignment #1 - Descriptive Statistics

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 is a resistant statistical measure, i.e., it is not sensitive to the extreme values that may occur. The median is a good measure of a typical value from a distribution. The median is also 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, may be interesting statistics. What would they mean for a qualitative or categorical variable?

The mean and median of a distribution may differ when a distribution is asymmetric but will be the same or similar for symmetric distributions. 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 standard deviation and variance take into account all of the measured values, i.e., they are calculated using all the values, thus they are not resistant statistics. The scale of the standard deviation is the same as that for the mean (whereas the variance is in squared units and may be harder to think about.) 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.

The range, maximum and minimum are measures of spread that reflect the extremes of the distribution. It is worth considering whether the highest and lowest value are true measured values that just happen to be very high or low, 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 a bit more descriptive and informative since they aren't just indicators of the most extreme values.

Histograms and Bar Charts

Histograms give a lot of information about the shape of the distribution for a numerical variable. You will want to notice whether the number of measurements of very low or very high values drops off quickly, i.e., whether the distribution is sharply peaked or fairly flat across all the measured values. For a categorical variable (such as commuting method) a bar chart shows the relative frequency of various categories. You will want to notice whether there are categories that are observed with much higher frequency than others or whether all of the categories are observed with about equal frequency.

The Data Set

The data set you used for this assignment is a population survey of Washington State. If our population of interest is households in Washington State then the data are great for answering questions that might interest us. 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 control or adjust for differences. To do this it would be necessary to collect data on households in other states. The unit of observation in this case is a household.

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

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

Households in the sample have on average 3.3 members and an average of \$65,444 in household income. The reported number of household members varies from 1 to 14. About 68% of households have between 1.65 people and 4.95 people. The figures show the distribution of household size and income. Less than 5 percent of the households had more than 6 members, and most households have a total income of \$100,000 to \$120,000 or less. About 68% of households in the sample have between no income and approximately \$220,000 of household income per year. (The standard deviation is very large, \$155607, and is attributable to some extremely high outlying income responses¹). In the future it would be of interest to look at the relationship between household size and income, perhaps using another measure of income, i.e., calculated per household member, to allow consideration of equity issues. Also, it is important with skewed distributions (such as those for income) to look at other measures of central tendency, such as the median.

NOTE TO STUDENTS: Below are copied the outputs from SPSS so that you will be able to see where the information reported above came from.² In future assignments we would like you to include graphics right in the body of the memo or text. So, here is simply an example for the first assignment to be sure everyone knows how to extract the information from the output as it appears.

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

² Further note that the data range for the histogram on household income was shrunk to better show the distribution of incomes closer to the middle values. In other words, some of the extreme outliers were eliminated.

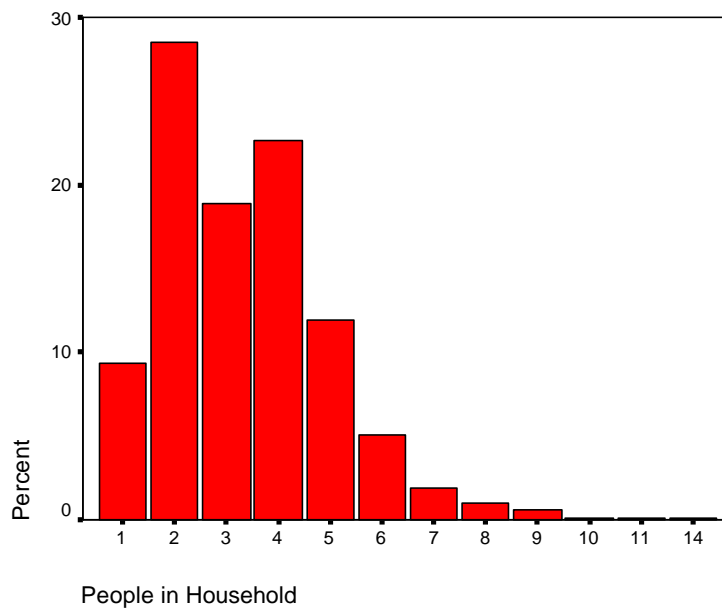
Variable 1 – Number of People in Household (possible descriptive output)

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	

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				

