ECO220Y1Y, Test #1, Prof. Murdock: SOLUTIONS

October 11, 2019, 9:10 – 11:00 am

NOTE: The parts of the solutions [in brackets] are extra explanations and are not required parts of your answer.

(1) (a) Choice **(C)**. The height of the bar from 1 to 1.5 is about 0.34 and the width of the bin is 0.5, which means about 3,060 tax rate changes =17,999*0.5*0.34. The height of the bar from 1.5 to 2 is about 0.12 and the width of the bin is 0.5, which means about 1,080 tax rate changes =17,999*0.5*0.12. Hence, the total number of tax rate changes from 1 to 2 is about 4,140 (=3,060+1,080).

(b) It is unimodal and *asymmetric* with very long, thin tails in both directions. The vast majority of municipal tax rate changes have been a modest increases of up to about one percentage point and only 7% have been decreases. However, a small fraction had considerable drops or considerable increases of several percentages points or more.

[This distribution is *not* symmetric, *not* Normal (Bell), and *not* well described as either positively skewed or negatively skewed. Implying any of those is a serious error. Of the standard terms we learned, only *unimodal* is applicable: the majority must be a context-specific description.]

(2) The table shows that even though City B has *lower* fare evasion rates on both streetcars and buses compared to City A, the overall fare evasion rate is *higher* in City B, which is a paradox. [The reason is that City B had a higher fraction of observations on streetcars, which generally are more prone to fare evasion.]

	City A			**City B**		
	Invalid payments	Observations	Evasion rate	Invalid payments	Observations	Evasion rate
Streetcars	750	3000	**0.250**	2400	12000	**0.200**
Buses	600	5000	**0.120**	400	4000	**0.100**
Overall	1350	8000	**0.169**	2800	16000	**0.175**

(3) In a 2019 survey, those strongly agreeing that vaccines are safe is a whopping 49 percentage points higher in South Asia (85%) versus Western Europe (36%), which is more than twice as high (136% higher).

(4) (a) Mean $= 3.52 = 0.042 * 1 + 0.049 * 2 + 0.254 * 3 + 0.655 * 4.$

(b) The 10th percentile is 3 because only 9.1% (4.2% + 4.9%) of the observations are a value of 1 or 2 so we are into the threes when we cross 10% of the data.

(c) Mean $= 0.919 = 0.081 * 0 + 0.919 * 1$ and s.d. $= \sqrt{0.081(0 - 0.919)^2 + 0.919(1 - 0.919)^2} = 0.273$

[This s.d. calculation ignores the slight degrees of freedom correction, but if you did it you'd still get the same answer with rounding: $\sqrt{(0.081(0 - 0.919)^2 + 0.919(1 - 0.919)^2) * \frac{1104}{1103}} = 0.273.$]

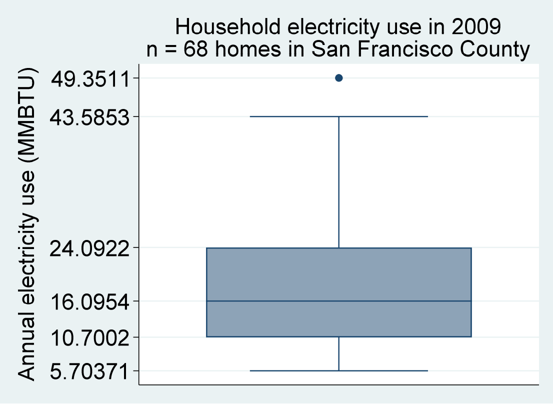
(d) Non-response bias is when there are systematic differences between the people who choose to respond to a survey versus everyone who was invited to participate in the survey. It is a type of non-sampling error. Table D.1 lets use compare some known features between those who responded versus everyone who was invited so we can look for any big differences. It appears that there are not substantial differences: the revenue distribution among responding firms is similar to all invited, as is profitability and industry. For example, mean revenue is very similar between all those invited versus those actually responding: 2.67 versus 2.74, respectively. Overall, Table D.1 suggests that we can rule out a serious non-response bias.

(e) A cross-tabulation and the associated bar chart, which provides a visual summary, would be the best way to summarize the relationship between these two nominal/categorical variables. We cannot use methods like the coefficient of correlation or a scatter plot because those are only appropriate for interval (quantitative) variables.

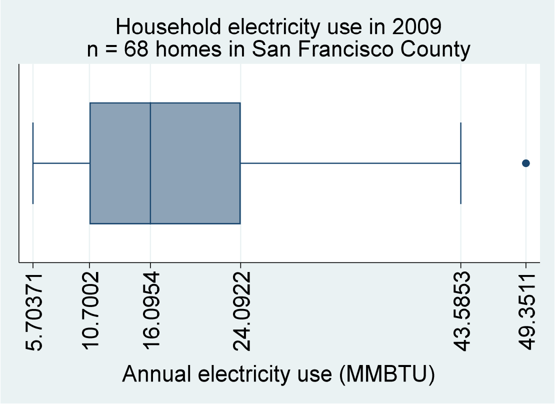
(5) The diagonal elements are variances and the off-diagonal elements are covariances. For example, the correlation between x and w is $r = \frac{s_{xy}}{s_x s_y} = \frac{88.1797}{\sqrt{70.8024}\sqrt{1038.88}} = 0.3251$. w and z have the strongest correlation: 0.6758.

```
             |        w        x        y        z
-------------+------------------------------------
           w |   1.0000
           x |   0.3251   1.0000
           y |   0.1711   0.1259   1.0000
           z |   0.6758   0.4332   0.1873   1.0000
```

(6) (a)



OR



(b) Both histograms have an extremely similar number of observations (6,844 is only 6 different from 6,838) and the number of bins depends on the number of observations [no matter which of the many formulas you use to give a suggested number of bins].

(c) The natural gas usage distribution is clearly bimodal because some homes are not connected to natural gas and hence use zero. However, we cannot see the bimodality in Histogram #1 because the range of the data is so large with the six outlier homes included. Hence, while the number of bins is the same, the bin width is much wider for Histogram #1, which lumps together homes that use no natural gas with those homes that use a relatively modest amount of natural gas.

(7) (a) How much does cumulative exposure to fine particulate matter air pollution (the x variable) affect the chances of getting dementia/Alzheimer's (the y variable) later in life?

(b) We'd *randomly assign* each person a level of long-term exposure to air pollution, with some getting air with high concentrations of PM2.5 and others getting much cleaner air. Then we'd observe their long-run health outcomes with respect to dementia. [Obviously this is not realistic: we cannot randomly assign people to live in a certain air quality for a long period of time. People, and especially the wealthy, would revolt, not to mention the ethical issues.]

(c) The locations where people live, and hence their exposure to PM2.5, is not random and is correlated with risk factors for dementia. For example, states with more poor people are more likely to have intense risk factors such as diabetes, smoking, and low education, and are also more likely to have a higher fraction living near highways, factories, and other more polluted locations where home and rental prices are cheaper. Hence, the positive correlation could reflect these other factors – called lurking/unobserved/omitted/confounding variables – rather than a real causal effect of air pollution on peoples' chances of suffering dementia in old age. In other words, 0.66 is very likely an upwardly biased estimate of the strength of the relationship between air pollution and dementia rates. [In other words, it suffers from endogeneity bias because PM2.5 is an endogenous x variable.]