Lab 1: Data Description and Visualization
KEY

Javier E. Flores

January 28, 2019

Total Possible Points: 34

Data

Q1) [2 pts] Suppose we want to use these data to learn about the studying habits of all students at Grinnell College. Are these data a population or a sample? Are there potential sources of bias? Explain.

A1) These data are a sample.[1 pts] The targeted population for this sample would be all students currently attending Grinnell College. Sources of bias include, but aren't limited to, selection bias, social desirability bias, and confirmation bias. Selection bias due to the fact that not everyone attending Grinnell is required to take this course; social desirability bias due to the fact that students may want to answer in such a way that they appear more studious (since I will be viewing responses); and confirmation bias due to the fact that not all students catalog their study hours and so remember them based on the kind of student they believe they are.[1 pts] Provided at least one source of bias is mentioned with reasonable explanation, appropriate credit should be given. If sources other than these are mentioned and backed up by a reasonable explanation, appropriate credit should be given.

Q2) [2 pts] What if we wanted to use these data to learn about the studying habits of the students enrolled in STA209-04 (this class)? Are the survey data a population or a sample? Is there bias? Explain.

A2) Since not all students answered the survey, these data are a sample.[1 pts] If, on the other hand, all students did complete the survey (as I had hoped) we would have responses from the entire population. Social desirability bias, or confirmation bias are are all potential sources of bias. Social desirability bias due to the fact that students may want to answer in such a way that they appear more studious (since I will be viewing responses), and confirmation bias due to the fact that not all students catalog their study hours and so remember them based on the kind of student they believe they are.[1 pts] Provided at least one source of bias is mentioned with reasonable explanation, appropriate credit should be given. If sources other than these are mentioned and backed up by a reasonable explanation, appropriate credit should be given.

Q3) [2 pts] Think back to the types of bias we previously discussed in class. Does the class survey exhibit any of these biases? If so, which? Explain.

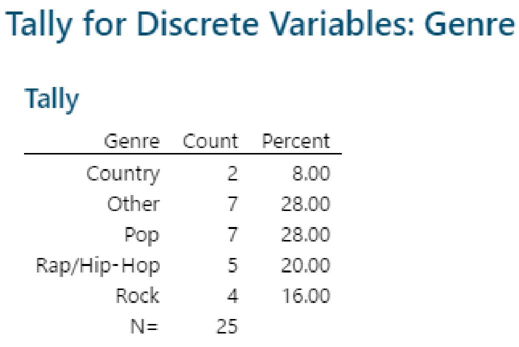
A3) The survey exhibits social desirability bias, confirmation bias, and leading questions.[1 pts] For explanations of the first two, see the above answers. An example of a leading question in the survey is ”Isn't Javi's dog, Mellow, the cutest?”. This question presupposes that Mellow is the cutest, and the answers even influence an affirmation of that claim.[1 pts] If a source of bias other than (and in addition to) these is listed and provided with reasonable explanation, credit should be given.

Categorical Variables

Single Categorical Variable

Q4) [1 pts] Create a frequency table for the survey question ”Which genre of music do you listen to most often?”. Include the result as a table in your write-up.

A4) Check for the requested table.[1 pts]



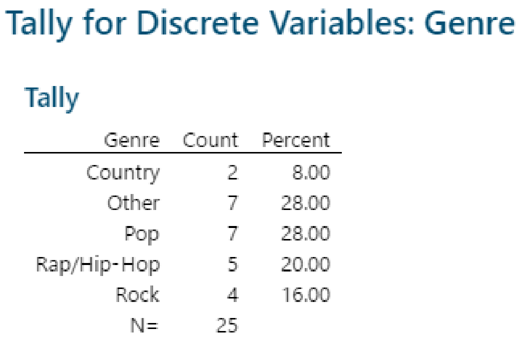
Tally for Discrete Variables: Genre

Tally

Genre	Count	Percent
Country	2	8.00
Other	7	28.00
Pop	7	28.00
Rap/Hip-Hop	5	20.00
Rock	4	16.00
N=	25	

Q5) [1 pts] Using the frequency table in Q4), compute the proportions for each category. Show your work and compare it to the appropriate Minitab output. Include the results as a table in your write-up. In order to reduce the length of your report, you may combine the tables for Q4) and Q5).

A5) Check for the requested table only.[1 pts]



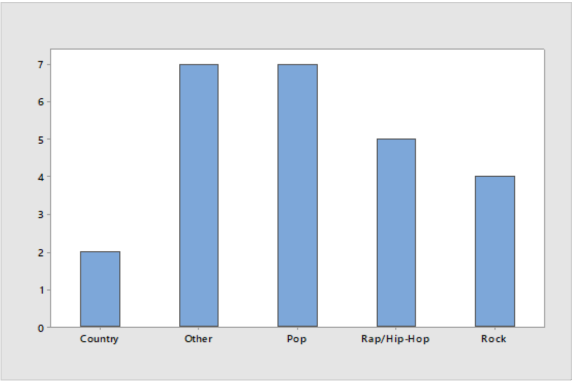
Tally for Discrete Variables: Genre

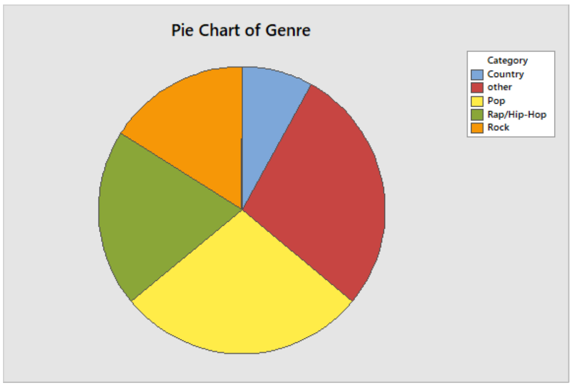
Tally

Genre	Count	Percent
Country	2	8.00
Other	7	28.00
Pop	7	28.00
Rap/Hip-Hop	5	20.00
Rock	4	16.00
N=	25	

Q6) [2 pts] Using Minitab, create both bar and pie charts for the survey question from Q4).

A6) Check for both of the requested figures.[2 pts]





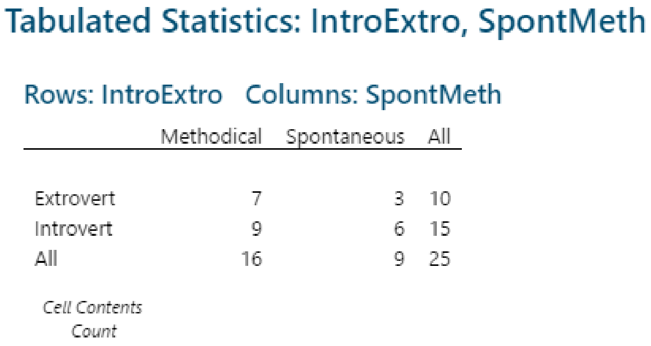
Q7) [1 pts] Which of these is more effective in communicating information? Or are they both equally effective? Explain your choice and rationale.

> **A7)** Any answer is correct here so long as provided rationale is reasonable.[1 pts] The bar chart may be more effective if your main interest is in directly visualizing relative frequencies, and the pie chart may be more effective if your main interest is in directly visualizing proportions. Both, however, sufficiently communicate both pieces of information. Proportions can be inferred using bar heights in bar charts, and frequencies can be inferred from pie charts provided that a total is given.

Two Categorical Variables

Q8) [1 pts] With Minitab, create a two-way frequency table using the question "Are you an introvert or extrovert?" as the row variable and the question "Are you spontaneous or methodical?" for the columns.

> **A8)** Check for the requested table.[1 pts]



Tabulated Statistics: IntroExtro, SpontMeth

Rows: IntroExtro   Columns: SpontMeth

```
               Methodical   Spontaneous   All
-----------------------------------------------

Extrovert               7             3    10
Introvert               9             6    15
All                    16             9    25

 Cell Contents
    Count
```

Q9) [2 pts] Compute all row-conditional and column-conditional probabilities. Show your work.

> **A9)** Check for the requested computations and work.[2 pts]
>
> - Row: $\pi_{met|ext} = 7/10$, $\pi_{spont|ext} = 3/10$, $\pi_{met|int} = 9/15$, $\pi_{spont|int} = 6/15$
> - Column: $\pi_{ext|met} = 7/16$, $\pi_{int|met} = 9/16$, $\pi_{ext|spont} = 3/9$, $\pi_{int|spont} = 6/9$

Q10) [2 pts] Which characteristic - spontaneous or methodical - is more likely to describe the extroverts in our class? Explain.

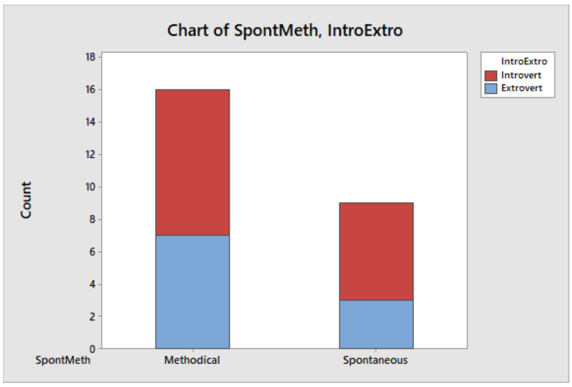
> **A10)** Here we are conditioning on students that are extroverts. Therefore, we compare the two row-conditional probabilities $\pi_{met|ext}$ and $\pi_{spont|ext}$. Since we found $\pi_{met|ext}$ to be larger, we would say that extroverts in our class are more likely to be methodical.[2 pts]

Q11) [2 pts] Are spontaneous individuals in our class more likely to be extroverts? Explain. (Think carefully before you answer!)

A11) In this question, we are conditioning on students that are spontaneous. Therefore, we compare the two column-conditional probabilities $\pi_{ext|spont}$ and $\pi_{int|spont}$. Since we found $\pi_{int|spont}$ to be larger, we would say that spontaneous individuals in our class are more likely to be introverts. [2 pts]

Q12) [1 pts] Create a stacked bar chart showing responses to "Are you an introvert or extrovert?" conditional upon being spontaneous or methodical.

A12) Check for the requested figure.[1 pts]

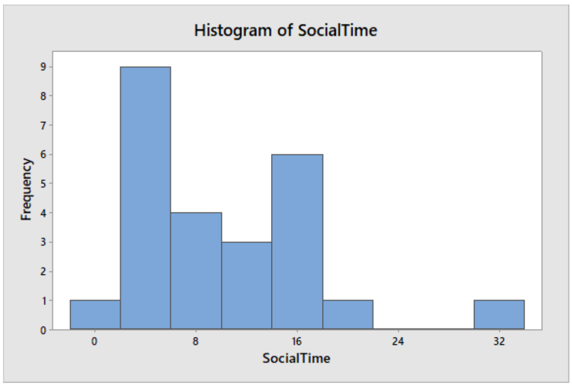


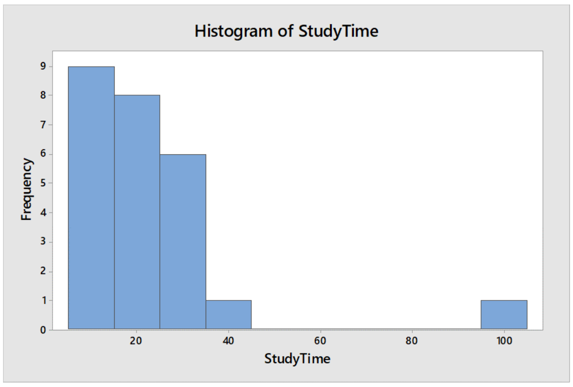
Quantitative Variables

Single Quantitative Variable

Q13) [2 pts] Create histograms of the responses to the questions "How much time (in hours) do you spend on social media each week?" and "How much time (in hours) do you spend studying each week?". Include both histograms as images in your write-up.

A13) Check for both of the requested figures.[2 pts]



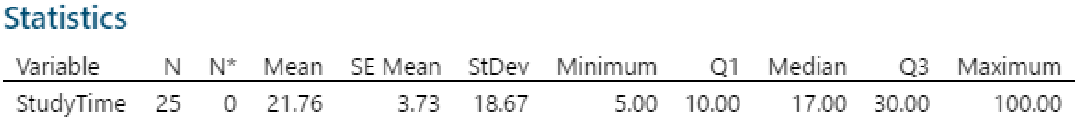


Q14) [2 pts] Describe the shape of each histogram created in Q13). What conclusions can you draw from these figures?

A14) For the SocialTime histogram, the histogram is roughly right skewed due to the data point at 32. Since there are two clear peaks, a more appropriate description for this histogram would be spiked or approximately bimodal. For the StudyTime histogram, the shape is more clearly right skewed. [1 pts] In visually comparing these two histograms, it would appear as if the variability in StudyTime is greater than SocialTime. Furthermore a greater proportion of students tend to spend 20 hours or more studying whereas the reverse is true in terms of time spent on social media.[1 pts]

Q15) [1 pts] Report the mean, median, standard deviation, and IQR of the responses to "How much time (in hours) do you spend studying each week?".

A15) Check for the requested statistics.[1 pts]



Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
StudyTime	25	0	21.76	3.73	18.67	5.00	10.00	17.00	30.00	100.00

IQR = 30-10 = 20

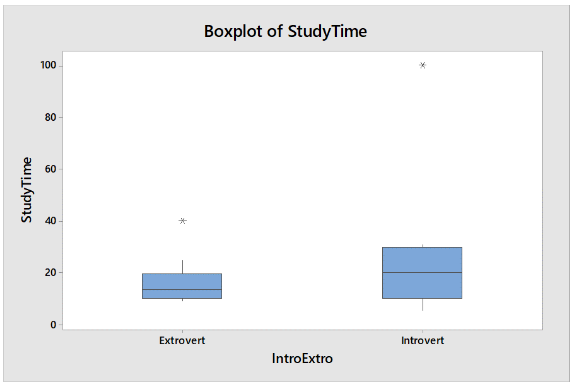
Q16) [2 pts] Suppose there was a student who (badly) lied about their study time in hopes of making a good impression, and reported studying 170 hours each week. Which of the statistics you reported in Q15) do you expect to change (if any)? If you think some statistics will change, explain how they would change and why.

A16) The mean and standard deviation would both change. The mean would increase since large values increase the mean, and the standard deviation would increase since it is a function of the mean (which increased). [2 pts] If only a change in mean was mentioned, give full credit.

(Visually) Bridging the Quantitative-Categorical Gap

Q17) [2 pts] Use boxplots to answer the question: "Do introverts spend more time studying than extroverts?" Include your plots along with a few sentences to explain your reasoning.

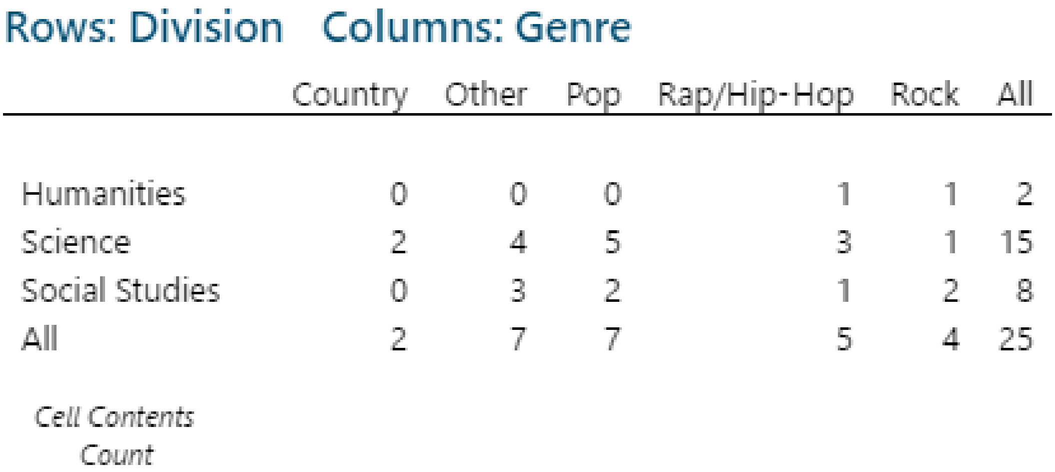
A17) Check for the appropriate figure and a reasonable answer/explanation. [2 pts]



It would appear that introverts spend more time studying than extroverts. The extrovert boxplot indicates that 75% of extroverts spend less than 20 hours studying. This means that only 25% spend more than 20 hours studying. In contrast, 50% of introverts spend more than 20 hours studying.

Q18) **[3 pts]** Use one or more of the techniques discussed in this lab to answer the question: "Which genre of music do Social Studies students listen to most often?" Include any relevant figures, and provide rationale for the chosen technique(s) and response. Do not exceed 5 sentences in your explanation.

A18) Check for appropriate figure(s) (e.g. pie chart), statistic (e.g. proportions), and a reasonable answer/explanation. [3 pts]



Rows: Division   Columns: Genre

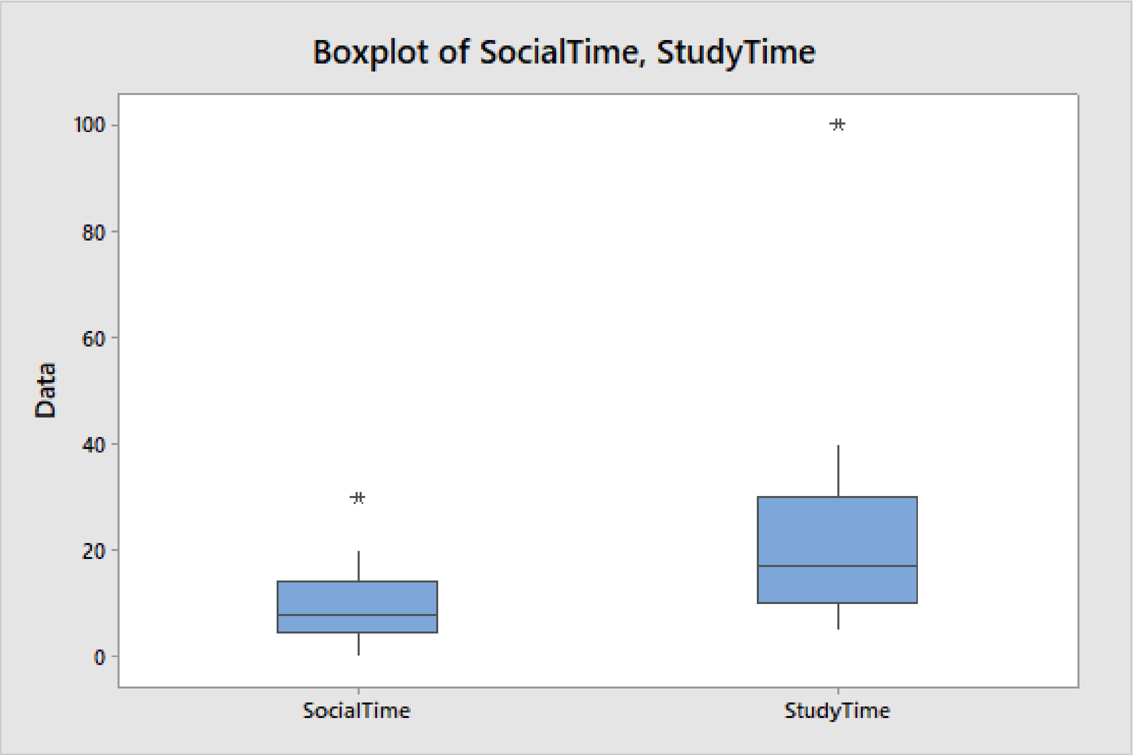
	Country	Other	Pop	Rap/Hip-Hop	Rock	All
Humanities	0	0	0	1	1	2
Science	2	4	5	3	1	15
Social Studies	0	3	2	1	2	8
All	2	7	7	5	4	25

Cell Contents
Count

Social Studies students listen to 'Other' music most often since 3/8 designated this as their survey response.

Q19) **[3 pts]** Use one or more of the techniques discussed in this lab to answer the question: "Do students spend more time on social media or studying?" Include any relevant figures, and provide rationale for the chosen technique(s) and response. Do not exceed 5 sentences in your explanation.

A19) Check for appropriate figure(s) (e.g. histograms), statistic (e.g. mean), and a reasonable answer/explanation. [3 pts]



Students appear to spend more time studying than they do on social media. 75% of students spend approximately 17 hours or less on social media. In contrast, 75% of students spend roughly 15 hours or more studying.