Alice Brawley-Chesworth
USP 634 – Spring 2018
Homework 1

1. **Descriptive analysis of your dataset.**

   My dataset contains only categorical variables (see Table 1). However, for the purpose of this assignment, I will use the "Question 1" and "Question 2" columns, since they are ordinal categorical variables, for the numeric variable assignments.

Comment [L1]: What are they?

	Sex ID	Ethnicity ID	Position ID	WorkGroupID	Year	Question1	Question2	…	Question16
1	F	W	U	ES	2001	4	4		4
2	M	W	U	NO	2001	2			
3	F	A	S	ES	2001	3	2		3
…									
1766	M	W	U	ES	2007	4	4		4

Table 1

Comment [L2]: Good to provide a snippet of your dataset.

SexID: F=female, M=male, NO=not specified
EthnicityID: A=Asian/Pacific Islander, AA=African American, H=Hispanic, NA=Native American, PI=Native Hawaiian or Pacific Islander, NS=Not Specified, W=White, T=Two or More Races
PositionID: C=Contract, M=Manager (Division or Group), N=Non-Rep/Non-Supervisory, NS=Not Specified, S=Non-Rep/Supervisory, U=Union Represented, O=Other
WorkGroupID: BS, DO, ES, IW, NO/NS = not specified, PP, WG, WS/PL (note: one work group – WS/PL - changed names over the course of the survey period, also early in the survey period NO was used when the question was not answered, later NS was used).
Year: 2001-2004, 2006-2007
Questions 1 – 16: 1=strongly disagree, 2=disagree, 3=agree, 4=strongly agree.

1) a– g. answers for Questions 1 & 2 can be found in Table 2.

		Question 1	Question 2
a.	Mean	3.061	2.693
b.	Mode	3	3
c.	Median	3	3
d.	Range	3	3
e.	Interquartile range	1	1
f.	Variance	0.5998	0.6020
g.	Standard deviation	0.7745	0.7759

Table 2

2) **Show with graphs and describe the distribution of each continuous variable. Which of these two variables resembles the normal distribution more closely? How can you tell?**

Using a histogram of the data, and a 'normal' curve superimposed, as in lab 2, we can visually inspect the two variables. And compare the fit to a normal distribution using a Q-Q plot:

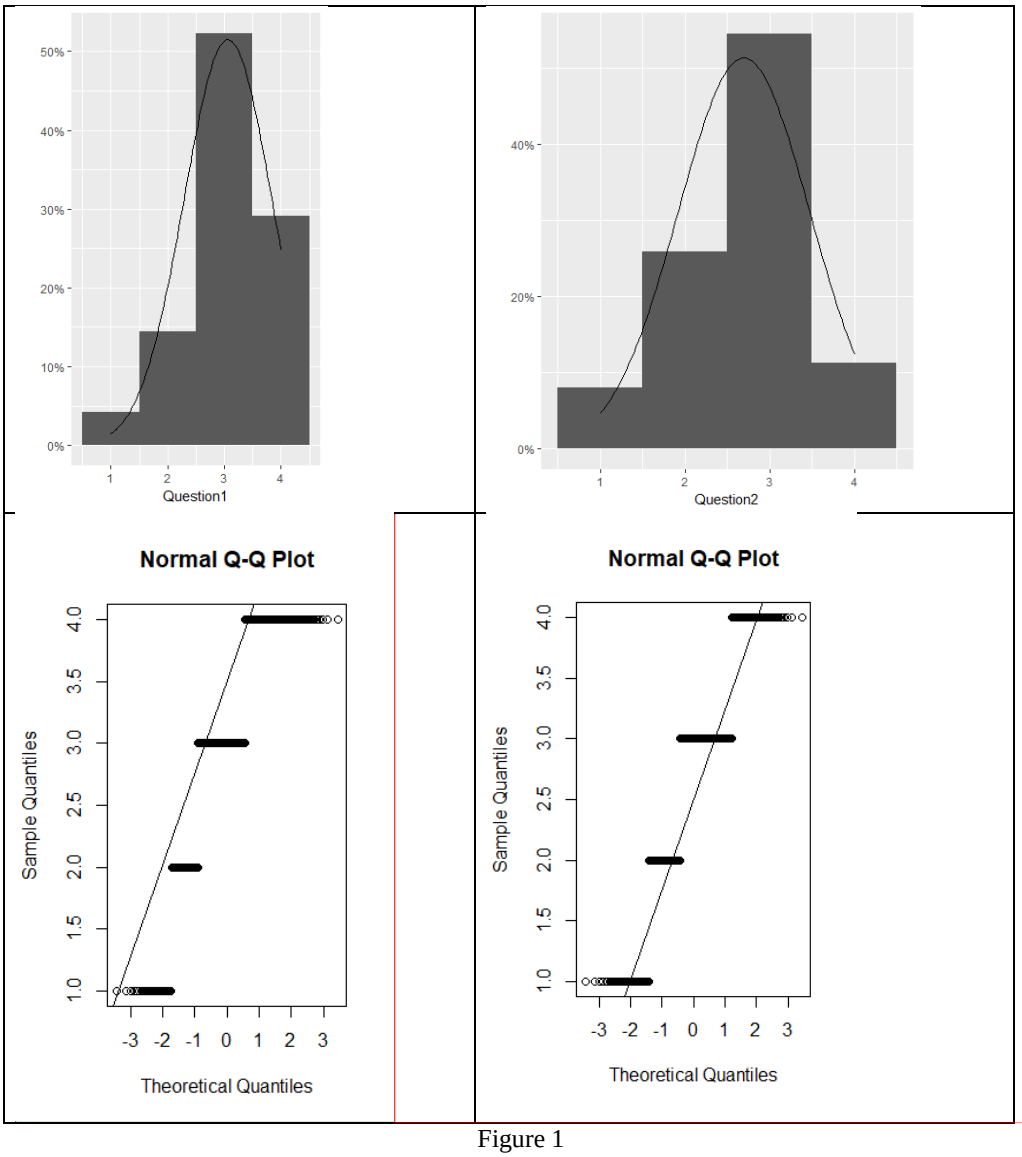

Figure 1

Comment [L3]: Good work using the qq-plot for checking normality.

Neither of these is very close to a normal distribution, but visually we can determine Question 2 appears to be a better fit.

3) **Show with appropriate graphs and describe the relationship between these two continuous variables. Are they dependent? If so, positively or negatively?**

Usually a scatter plot of the two continuous variables can be useful for visually determining if they are independent. However, since these are actually categorical variables that are being treated as numeric, the graph does not help much:

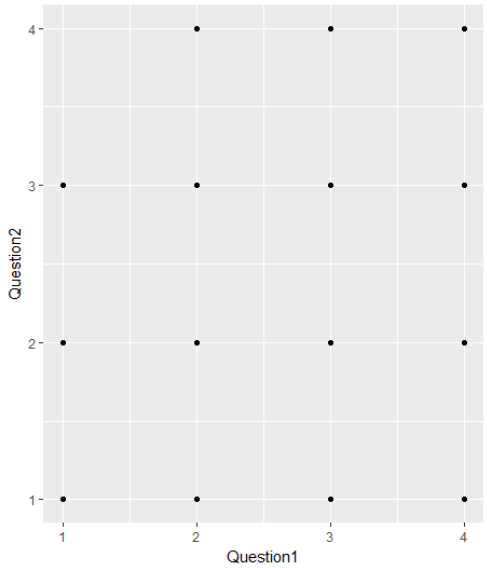

Figure 2

Instead, we will use a hypothetical example to examine independence. We will see whether the probability of getting an answer of "3" in Question 2 correlates with the probability of getting an answer of "3" in Question 1. (Note: Question 1 is "I feel safe discussing an issue or problem with my co-workers" and Question 2 is "Conflicts are worked out", so what we're testing is whether the probability that someone agreed with "Conflicts are worked out" if they also agreed with "I feel safe discussing an issue or problem with my co-workers").

For this, we need to find the probability of a 3 in Question 2
$P(Q2=3) = 896/1766 = 0.5074$
And the Probability of getting a 3 in Question two, given there was a 3 in Question 1
$P(Q2=3 \mid Q1=3) = 582/1766 = 0.3296$

These two values are not the same, so the two variables are not dependent.

Comment [L4]: Correct. A trick when your "numeric" variable in your scatter plot has few unique values is to use geom_jitter in the place of geom_point. (http://ggplot2.tidyverse.org/reference/geom_jitter.html)

Comment [L5]: I think this means they are dependent. If they are independent, they conditional probability is the same as the unconditional one.

4) **For each of the 2 categorical variables, show with appropriate graphs and tables, and describe the distribution.**

For this I will use the two categorical variables: SexID and WorkGroupID.
For the SexID variable (Figure 3), 61.4% of respondents were male, 38.0% were female, and 0.62% did not respond to this question. For the WorkGroupID variable, the data first needs some cleaning; PL and WS will be merged to become WS (the new name of the work group), and NS and NO are merged into NO (not specified). The percentages can be seen in Table 3; For this question, 31.6% of respondents did not respond to the question, the largest value for the WorkGroup question. The most frequent positive responses were ES (24.7%) and WG (16.0%). See Table 3.

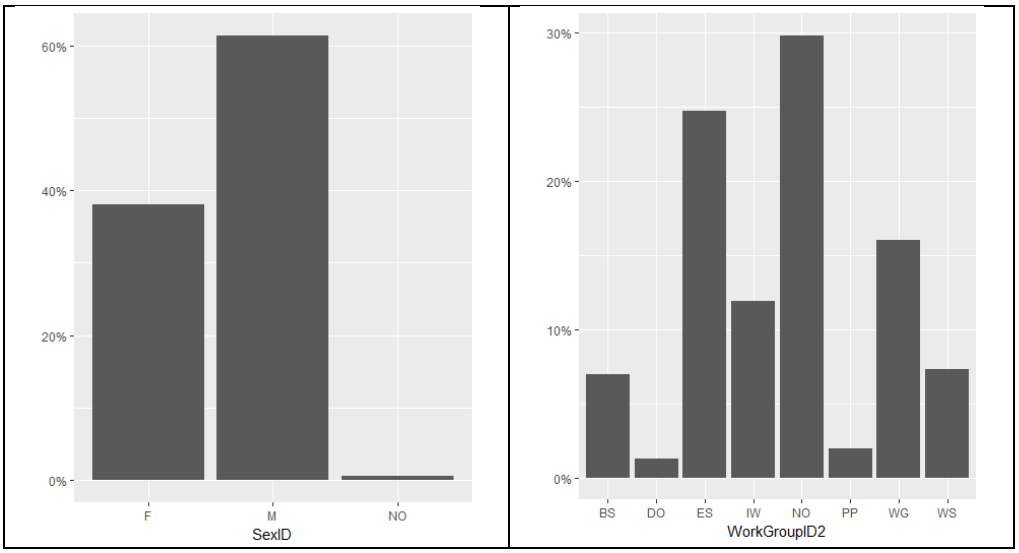

Figure 3

	BS	DO	ES	IW	NO (includes NS)	PP	WG	WS (includes PL)
Count	123	22	437	210	527	35	283	129
Percent	6.96	1.25	24.7	11.9	31.6	1.98	16.0	7.74

Table 3

5) **Show with appropriate graphs and describe the relationship between the 2 categorical variables. Are they dependent?**

Figure 4 shows a bar chart of the two variables. From this chart we can see that the two are not dependent.

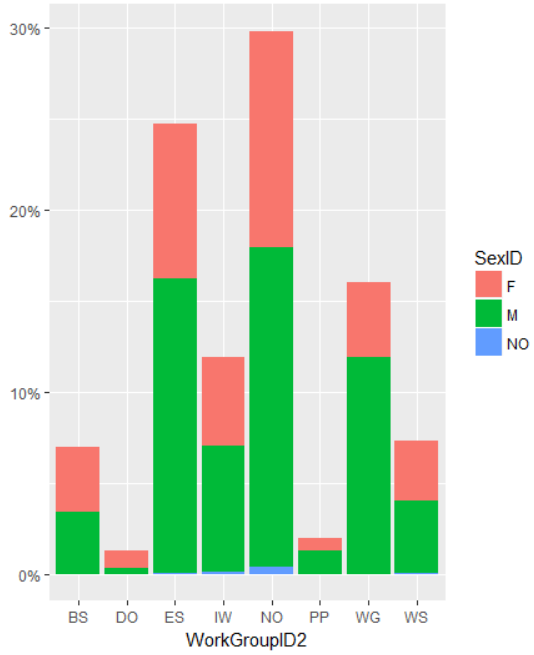

Figure 4

Comment [L6]: One possible improvement: for detecting relationship between two categorical variable, use geom_bar(position="fill"). See http://ggplot2.tidyverse.org/reference/position_stack.html for more explanation.

6) Select one continuous variable and one categorical variable, show with an appropriate graph and describe the relationship between the two variables.

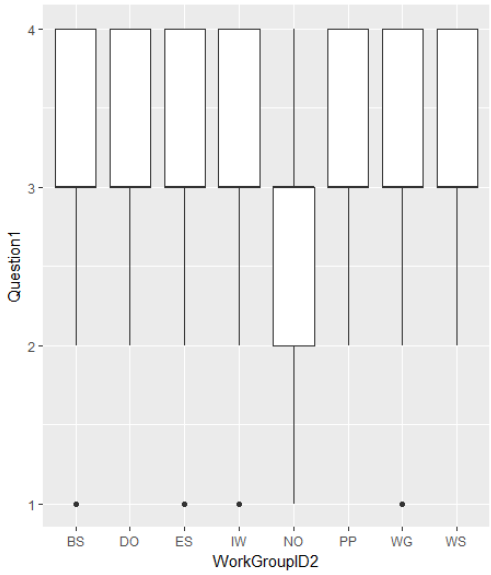

Figure 5

The box plot in Figure 5 shows the relationship between Work Group and answers to Question 1. It is interesting to see that respondents who declined to specify their Work Group were more likely to disagree with the statement "I feel safe discussing an issue or problem with my co-workers".

7) **Among the exercises above, what do you think is the most interesting relationship you find between a pair of variables in your data? Why?**

I find the difference in answers to Question 1 between those who positively answer the Work Group question and those who do not to be the most interesting. This suggests that those employees who filled out the survey who did not feel comfortable providing information that could potentially identify them among their co-workers are also the same ones who are less likely to disagree that they felt safe discussing issues with their co-workers. Although this is not counter-intuitive, it will need to be taken into account in my further work trying to determine if survey-responses vary between work groups, especially in light of the high number of respondents who did not indicate their work group. Further investigations into how this may have changed over time will need to be made before any tentative conclusions can be drawn.

2. **Data collected at elementary schools in DeKalb County, GA suggest that each year roughly 25% of students miss exactly one day of school, 15% miss 2 days, and 28% miss 3 or more days due to sickness.**

   a) **What is the probability that a student chosen at random doesn't miss any days of school due to sickness this year?**

   For every 100 students, 25+15+28 = 68 will miss at least one day of school, the number that will miss no days is 100 - 68 = 32. Therefore,

   P (no days missed) = 32/100 = 0.32

   b) **What is the probability that a student chosen at random misses no more than one day?**

   Missing no more than one day equals missing zero days (32) plus missing one day (25) = 57.

   P (no more than one day missed) = 57/100 = 0.57

   c) **What is the probability that a student chosen at random misses at least one day?**

   Missing at least one day is the disjoint is missing no days. Therefore,

   P (at least one day missed) = 1 – P (no days missed) = 1 – 0.32 = 0.68

d) If a parent has two kids at a DeKalb County elementary school, what is the probability that neither kid will miss any school? Note any assumption you must make to answer this question.

For this problem, independence will be assumed. While we know that this is not strictly true, in the same way that removing one card from a deck changes the probabilities of which card will be drawn next, for a large population of pupils, the change is probability based on knowledge of the status of one student does can be assumed to be inconsequential. This assumption is reasonable, given that DeKalb County had about 178,000 people under age 18 in 2017 (https://www.census.gov/quickfacts/fact/table/dekalbcountygeorgia/PST045217). In addition, we are assuming that one child being out sick does not influence the health of the second child, something any parent knows is not actually true since close contact with an infected person increases the chances of the second person getting sick.

Using these assumptions, we can calculate the Probability that 2 children will both not miss any school as:

P (no days missed) * P (no days missed) = 0.10

e) If a parent has two kids at a DeKalb County elementary school, what is the probability that both kids will miss some school, i.e. at least one day? Note any assumptions you make.

Using the same assumptions given in part d) of this question, the calculation is:

P (at least one day missed) * P (at least one day missed) = 0.68 * 0.68 = 0.46

f) If you made an assumption in part (d) or (e), do you think it was reasonable? If you didn't make any assumptions, double check your earlier answers.

See the answer to part d). The assumption, based on population numbers, that knowledge of the status of one child will not significantly impact the probability of a second child missing school is reasonable. However, the second assumption, that a sick child will not infect a sibling, is less supportable. Research has shown a correlation between sibling health status and school sick days, though there is a larger correlation with other, nonfamiliar influences.

Source: Wilcox-Gök, V. L. (1983). Sibling data and the family background influence on child health. *Medical care*, 630-638.

Comment [L7]: Nice!

3. The average daily temperature in Orangetown is approximately normally distributed with a mean of 75F and a standard deviation of 15F. What is the probability for the average daily temperature to be below 55F or above 90F, which may damage the orange crop?

55 degrees is 20F below the mean, or 1.33 standard deviations. Using pnorm in R, we get 0.0912 for probability of being below 55 degrees, and 0.8413 for the probability of the temperature being above 90 degrees (or 0.1587 of it being above). Therefore,

P (below 55 or above 90) = P(below 55) + P(above 90) = 0.0912 + 0.1587 = 0.2499

Comment [LW8]: +

How many days in a year do the average daily temperature stay in the range of 55-90F?

We would expect the temperature to stay in the range of 55-90F 75% of the time, so 365 days/year * 0.75 = 274 days per year.

4. **Inferring the direction and existence of causal relationships from observational data can be plagued by selection bias, reverse causality, and confounding variables (a third variable or a number of other variables, influence both explanatory and response variables). The following empirical patterns have been cited in press reports as potential evidence of causal relationships.**

    - **Oakland is considering a Fresh Food Financing program that incentivizes grocery stores to locate in East Oakland. This program is based on studies showing that residents of neighborhoods without stores selling fresh foods have an unhealthy diet.**

    Comment [L9]: From this description, it seems there may also be selection bias.

    - **Two percent of residents in Fresno, CA bike to work while eight percent bike in Berkeley. Berkeley has 50 more miles of bike lanes on their roads than Fresno. Therefore, if Fresno were to add more bike lanes its bike ridership would increase.**

    - **A recent study in Minneapolis found that people who live in neighborhoods where the majority of houses have porches are more likely to talk to their neighbors at least once a week in comparison with people who live in neighborhoods where there are few porches. To encourage social cohesion in neighborhoods, Minneapolis is therefore considering a new grant program to help people add porches to their houses.**

    All three empirical patterns are seen in observational (non-experimental) data. Can you apply any of the criticisms of non-experimental empirical results to these three examples? If these criticisms were true, how do they alter interpretation of these patterns?

    In general, there are three potential explanations for a correlation between two variables in observational data. 1. A causes B; 2. B causes A; 3. Both A and B are caused by a third factor, C.

For the Oakland example, the three options are: 1. The lack of fresh food causes an unhealthy diet, 2. An unhealth diet among residents prevents establishment of fresh food stores, or 3. A third variable, such as poverty, or the residents working more than 40 hours per week, therefore lacking the time or energy to prepare fresh foods, cause both unhealthy diets and the inability of fresh food stores to survive in the neighborhoods. If either 2 or 3 are true, placing fresh food stores in the neighborhood will not help improve diets and will be a bad economic investment as well.

Similarly for the Fresno/Berkeley example, 1. Presence of bike lanes may increase bike commuting, 2. Bike commuters cause more bike lanes to be built, or 3. Some other, third variable, such as median household income, or environmental orientation, could explain both bike commuting and bike lane construction in a community. If Oakland spends resources on bike lanes and 2 or 3 are true, bike commuting will not increase and public funds will have been wasted on useless infrastructure that could have been used for a other public goods.

Finally, we have the same reasoning in the Minneapolis example: Either 1. Front porches facilitate social cohesion, 2. People who value social cohesion move to neighborhoods with front porches, or 3. Something else, such as median household income, aesthetic sensibilities, lead both to social cohesion and a preference for front porches. Again, just as in the other two examples, the authorities assume that 1 is true, but if 2 or 3 are true, then the goals will not be met and resources will be wasted.