Survey Analysis for Student Minds

CSRN Data Team

Introduction

This report will examine the data collected in the Student Minds survey, detail how we cleaned the data, explain why we made the decisions we made and briefly analyse the results. The aim of this report is to answer three questions:

1. Were there any correlations between concern for well-being support, concern for social life and concern about loneliness, each with every other question?
2. Were there any associations between any of the respondent's demographics and their answers?
3. For each demographic was Bristol performing better or worse than other universities?

Before we answer these questions we need to consider the data.

The data

The data came from a survey produced by the Student Minds team. It contained 11 questions asking how helpful respondents found things such as university, parents etc, followed by 10 questions asking how concerned students were about other sources of support. There were also 4 free-text questions which this report will not go into. Finally there were 7 demographics questions. See Table 6 in the appendix which contains the questions with their respective percent missing.

Cleaning the data

First we cleaned the *universities* column. For this there were actually two columns caused by a drop-down menu for more common universities followed by another question for if their university wasn't mentioned. To clean this we created a new column that contained the results (converted to lower-case and spaces removed) of the second column. Then filled the blanks with the results of the first column. Unfortunately many students entered their university name in a non-standard format so we looked through all unique responses to that question and replaced all the non–standard options with a standard (we chose all lower case with just the university name so "Uni of exeter" became "exeter"). After doing this we realised that our respondents were very biased towards University of Bristol students - Bristol had 86 respondents, the next highest was Nottingham with 27 then Oxford with 19. We also had 12 universities with only 1 respondent so we created a new column that was true if the respondent went to Bristol and false if they didn't. Differences were then explored in our answer to question 3.

The next column to look at was *Gender*. Here we noticed the option 'prefer not to say' which appeared across many columns. We replaced all of them with NumPy NaN values so they would be ignored when we computed our statistics. Unfortunately we also had to do this for Non-Binary responses as we only had 2 and couldn't combine them with any other column.

Next we looked at the *sexuality* column combined with the column where respondents could indicate whether their gender matched the gender they were assigned at birth. Many of the results here were not frequent enough to be used for association analysis so we created a new column that was true if the respondent was LGBT and false if they were not.

We applied the same technique for the same reasons to the *ethnicity* column, creating a new column that was true if the respondent were BAME and false if not.

Unfortunately we had to ignore the columns detailing the respondent's *year* because it was too hard to clean given the time restrictions and *source* because we didn't believe it was as relevant as other columns.

See Table 1 for the final demographics features.

Table 1: Possible values for each demographic

Demographic	Values
Gender	Woman, Man
Student Status	Home, EU, International
is LGBT	True, False
is BAME	True, False
is Bristol	True, False

For the question columns, respondents could respond with numbers 1-5 indicating how helpful / concerned they were with that particular item. For the association analysis this was converted into groups 1-2, 3 and 4-5 to ensure all contingency table cells had expected values greater than 5.

Now the data is cleaned we can move on to answering the questions.

Correlations

First let's answer question 1:

> 'Were there any correlations between concern for well-being support, concern for social life and concern about loneliness, each with every other question?'

The responses followed a Likert scale (e.g. very concerned, mildly concerned, not concerned, etc). In order to look at correlation between these questions we used Spearman's Rank Correlation Coefficient r_s[2]. This is calculated by first ranking the data in increasing order from 1 to n and then using the PMCC defined as

$$r_s = \frac{cov(rg_x, rg_y)}{\sigma_{rg_x}\sigma_{rg_y}}$$

on the ranked data rg_x, rg_y[3]. The result is a value between -1 and 1 where $r_s = 1$ indicates perfect positive correlation and $r_s = -1$ indicates perfect negative correlation. We then calculated r_s for each pair of questions in our data.

Under the assumption that $r_s = 0$ we have that

$$T = r_s\sqrt{\frac{n-2}{1-r_s^2}}$$

is distributed according to a t-distribution with $n-2$ degrees of freedom. Therefore we can calculate the probability that we see our question answers, given that the question answers are not correlated. This will be the p-value and so a very low p-value would indicate that there is evidence to suggest that the data is correlated.

Results

Table 2 shows the top 10 most significant p-values grouped by the identified fields.

Table 2: Question correlations

Question 1	Question 2	r_s	p-value
C: loneliness	C: social life	0.58	7.2e-24
C: loneliness	C: feeling motivated	0.44	2.3e-13
C: loneliness	C: financial stress	0.34	3.4e-08
C: loneliness	C: physical health	0.31	5.3e-07
C: well-being support	C: academic support	0.57	9.2e-23
C: well-being support	C: feeling motivated	0.40	4.3e-11
C: well-being support	C: loneliness	0.40	5.4e-11
C: well-being support	C: career prospects	0.30	1.9e-06
C: well-being support	C: physical health	0.29	2e-06
C: social life	C: feeling motivated	0.33	1.1e-07

Analysis

The p-values appear extremely small in this table for what seems like r_s values which are not close to ± 1. However, since the dataset was quite large, the degrees of freedom of the t-distribution are very large. This means that the p-values are made much more extreme since having $r_s > 0.2$ for $n = 251$ is very unlikely. However, it is important to put these results in the context of being opinionated metrics and so the correlation could be caused by behaviour when responding to a survey rather than being representative of the actual correlations. Therefore, the p-values are much more useful in ranking how likely these questions are correlated rather than giving a definite probability of error.

Associations

Now we will tackle question 2:

> 'Were there any associations between any of the respondent's demographics and their answers?'

We're trying to find which questions have an association with a particular demographic. What this means is we're trying to find the questions and demographics for which knowing the value of one tells us something about the other.

We can test for this using a χ^2 test. Key assumptions that must be met for the χ^2 test are that all the cells of the contingency table must have expected values greater than 5 and must be frequencies not percentages. To compute the χ^2 statistic we first need to compute the expected frequencies for each cell. For a cell in the ith row and jth column, it's expected frequency is

$$E_{ij} = \frac{R_i \times C_j}{G}$$

where R_i is the total of the ith row, C_j is the total of the jth column and G is the grand total of the table. Then we compute the χ^2 statistic as

$$\chi^2_\nu = \sum_{i,j} \frac{(E_{ij}-O_{ij})^2}{E_{ij}}$$

where O_{ij} is the observed frequency of the i,jth cell and ν is the number of degrees of freedom[3]. Then we can compute the p-value from the χ^2_ν distribution under the assumption that there is no association ($\chi^2_\nu = 0$). Note that

$$\nu = (\text{number of rows} - 1) \times (\text{number of columns} - 1).$$

We can calculate this p-value for each combination of demographic and non-demographic questions then report all those which have a p-value less than 10%.

Results

Note in Table 3 'C' means concerned, 'H' means helpful and 'MH' means mental health.

Table 3: Association results

Question	Dem.	p-value
C: mental health	Gender	<0.001
C: well-being support	Gender	0.007
C: feeling motivated	Gender	0.011
C: academic support	Gender	0.017
C: financial stress	Gender	0.021
H: doctor	Gender	0.026
C: career prospects	Gender	0.029
C: mental health	is LGBT	0.032
H: parents	is BAME	0.068
C: academic support	is Bristol	0.073
C: well-being support	is BAME	0.080
H: student MH supporters	is Bristol	0.085
C: career prospects	is BAME	0.093
H: digital services	is Bristol	0.096
C: loneliness	Gender	0.097

Analysis

Immediately the first thing that jumps out is *gender* is the most common associated demographic. Additionally the *concerned* questions were significantly more associated than the *helpful* questions. This might be due to *concerned* questions receiving more responses as many respondents might not have used those specific mental health services.

In Table 4 we see each question with a comment. The comment is based on the characteristic that had the most significant direction so if it says 'Women are more concerned' then it means more women were in the 4-5 category than the 3 or 1-2 category and this disparity was greater than it was for the men.

Table 4: Association comments

Question	Comment
C: mental health	Women more concerned
C: well-being support	Women more concerned
C: feeling motivated	Women more concerned
C: academic support	Women more concerned
C: financial stress	Women more concerned
H: doctor	Women less concerned
C: career prospects	Men more concerned
C: mental health	LGBT more concerned
H: parents	Non-BAME more helped
C: academic support	Non-Bristol more concerned
C: well-being support	Non-BAME more concerned
H: student MH supporters	Bristol less helped
C: career prospects	Non-BAME more concerned
H: digital services	Bristol less helped
C: loneliness	Women more concerned

Bristol vs Non-Bristol

Finally we will turn to question 3.

'For each demographic was Bristol performing better or worse than other universities?'

The survey data obtained for Student Minds contained feedback from 254 students nationwide. Of these students 84 attended the University of Bristol. By considering the the remaining 170 student to represent a sample population, confidence intervals can be used to compare the feedback from Bristol Students to other UK universities.

Confidence Intervals

When looking at sample data it is important to note that the sample mean is likely to differ from the true population mean due to the randomness of the sample data. A confidence interval can be calculated for the observed data. This confidence interval is a range for which we can state with a certain percentage confidence will contain the true population mean.

So for this project we calculated a 95% confidence interval from the non-Bristol universities then checked to see if the mean from the Bristol results fell within this interval. The confidence interval bounds are defined as

$$CI = \bar{x} \pm z\frac{\sigma}{\sqrt{n}}$$

where $\bar{x}$ is the sample mean, σ is the standard deviation of the sample, n is the sample size and z is 'z-score' which for a 95% confidence interval is 1.96[1].

Results

We tested questions that could be directly affected by whether a student attended Bristol or not. We also broke down the question on how students felt the university helped into BAME and LGBT groups. Note in Table 5, CI L is the lower bound for the confidence interval and CI U is the upper bound.

Table 5: Confidence Intervals

Question	CI L	CI U	Bristol
H: academic/personal tutor	2.30	2.77	2.72
H: student well-being advisor	2.15	2.69	2.33
C: mental health	3.78	4.13	3.87
C: financial stress	2.80	3.25	3.12
H: student MH supporters	2.61	3.39	2.17

Analysis

Clearly, as shown in Table 5, the only question where Bristol did not lie in its respective confidence interval was the 'helpfulness of student-led mental health supporters'. Here the mean of the feedback given by University of Bristol students was significantly lower than the lower bound of the confidence interval. The data supplied through the survey therefore suggests that Bristol is under performing in the helpfulness of student-led mental health services.

References

[1] Avijit Hazra. Using the confidence interval confidently. *Journal of Thoracic Disease*, 9(10):4124–4129, October 2017.

[2] Geoff Norman. Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, 15(5):625–632, December 2010.

[3] T. D. V Swinscow and Michael J Campbell. *Statistics at square one*. BMJ, London, 2002. OCLC: 48468226.

Appendix

Table 6: Features and percent missing

Feature	Percent missing
H: university	3.11%
H: parents	3.89%
H: wider family	14.00%
H: partner	43.58%
H: friends	3.50%
H: doctor	42.41%
H: student well-being advisor	47.08%
H: academic/personal tutor	20.62%
H: MH professionals	57.59%
H: student MH supporters	63.04%
H: digital services	61.87%
C: academic support	10.51%
C: well-being support	8.17%
C: feeling motivated	7.39%
C: loneliness	7.00%
C: social life	7.00%
C: financial stress	8.17%
C: university economic sustainability	10.12%
C: career prospects	9.34%
C: physical health	8.17%
C: mental health	8.95%
Uni1	19.84%
Uni2	81.71%
Year	3.11%
Gender	3.11%
Gender alignment	3.11%
Sexuality	4.28%
Student status	3.50%
Ethnicity	3.11%