01_test

Cal Chengqi Fang

2/19/2022

1. Sample Description

1.1 Sampling Strategy

The dataset we will use in this project is the outcome of the NCSS-SRH survey project. This project used a combination of multi-stage random sampling method and snowball convenience sampling method to collect the data.

In the first stage, researchers built the sampling frame based on the distribution of higher education schools in the eastern, central, and western regions, referring to the directory of higher education institutions published by the Ministry of Education of China in June 2018. Due to the severely uneven distribution of higher education institutions in China at the provincial (administrative region level equal to states in America) level (e.g., there are only seven colleges and universities in the whole Tibetan region), there are difficulties in sampling at the provincial level, and the stratification of the sample colleges and universities in the first sampling stage was performed using the East, Central and West divisions based on administrative divisions provided by the National Bureau of Statistics.

In the second stage, colleges offering associate's or bachelor's degree were selected from each region. Selected colleges offering bachelor's degrees were then divided into four levels in each region: first-class universities, universities of first-class disciplines, general undergraduate colleges and universities, and private undergraduate colleges and universities. Institutions offering associate's degrees were divided into three levels: key associate's colleges and universities, general associate's colleges and universities, and private associate's colleges and universities. Corresponding number of universities are selected according to the proportion of each level in the overall, with an appropriate inclination to the colleges and universities with more geographically diverse student source, taking into account the capability of the survey implementation organization, China Youth Network. A total of 241 sample colleges and universities are selected. The purpose of sampling at school level in this way is to ensure that the social networks in the study have nationwide first-level dissemination points, and to ensure that college students in 31 provinces and regions nationwide have access to the questionnaire through social media and can be included in the sample (in fact, college students in 34 provincial-level administrative regions, including Hong Kong, Macao and Taiwan, are included in the snowballing process). However, the samples from Hong Kong, Macao and Taiwan were suggested to be excluded in the analysis stage because of the significant differences in social background, cultural environment and education system between these samples and those from mainland China.

The 241 sample universities were used as the starting point for snowballing to sample college students enrolled in higher education institutions nationwide. All samples were *volunteer samples* recruited in the form of convenience sampling, with students deciding whether or not to participate in the survey and being able to withdraw at any time during the process after participation, with full disclosure of survey ethics given in the guidelines.

Referring to the literature review, we could see that most of the similar existing researches on youth sexual knowledge, attitudes, and behavior conducted in other Asian countries used random samples. However, except some researches done in Thailand benefiting from the nationwide military mandate lottery, the random

samples used in existing researches mostly were generated from a relatively small pool, for instance, 12 schools (Hedayati-Moghaddam et al., 2015), 1 community (Jaya & Hindin, 2009), or 3 factories (Tang et al., 2011). In that sense, although the random selection of our data might be limited due to adopting snowball sampling, the comparatively large size might be able to compensate and ensure some representativeness.

1.2 Sample Structure

Because of the limited size of graduate students sample, I excluded all graduate student samples here. (P.S. 0 in the variable gender refers to female students while 1 refers to male students.)

1.2.1 Enrollment Year

	2011	2012	2013	2014	2015	2016	2017	2018	2019
0	62	59	136	270	693	2341	3880	7931	19103
1	59	62	90	151	342	1103	1680	3434	11102

As above shows, for both females and males, this dataset has more sample of students enrolled in 2019, who are the first-year college students at the time of investigation. There doesn't seem to exist much variation between the distributions of enrollment year of males and females.

1.2.2 City Level

	1	2	3	4	5	6
0	5698	8310	6953	6773	6207	534
1	2763	5959	4024	2992	2169	116

The female sample distributed comparatively even among level 1 to 5 cities, while the male samples seem more concentrated in the level 2 and level 3 cities.

1.2.3 School Tier

	associate1	associate2	associate3	bachelor1	bachelor2	bachelor3	bachelor4
0	2306	7137	1065	2123	2214	14371	5259
1	2993	3000	918	1530	976	6336	2270

There are more female samples in every school tier level except 1st level associate's degree institutions.

1.2.4 chisq-test on sample balance

```
##
##  Pearson's Chi-squared test
##
## data:  gp_describe$gender and gp_describe$syear
## X-squared = 244.75, df = 8, p-value < 2.2e-16
```

```
##
##  Pearson's Chi-squared test
##
## data:  gp_describe$gender and gp_describe$scityl
## X-squared = 788.12, df = 5, p-value < 2.2e-16
```

```
##
##  Pearson's Chi-squared test
##
## data:  gp_describe$gender and gp_describe$slevel
## X-squared = 1678.6, df = 6, p-value < 2.2e-16
```

All of the results of the three chi-square tests on enrollment year, city level, and school tier over gender are statistically significant, which means we will have to re-weight the sample before we further analyze it.

2. Draft Tests on Penerative Sex

2.1 Cross-tables and chisq-tests

2.1.1 City Level

12	13	14	15	16	17	18	19	20	21	22	23	24	25
0	0.01	0.01	0.03	0.07	0.11	0.27	0.20	0.17	0.09	0.04	0.01	0.00	0.00
0	0.00	0.01	0.04	0.07	0.14	0.29	0.18	0.15	0.07	0.03	0.01	0.00	0.00
0	0.00	0.01	0.02	0.06	0.12	0.28	0.19	0.15	0.08	0.05	0.01	0.01	0.00
0	0.00	0.02	0.04	0.06	0.12	0.30	0.19	0.13	0.08	0.04	0.02	0.01	0.01
0	0.00	0.01	0.02	0.06	0.08	0.27	0.19	0.22	0.07	0.05	0.03	0.01	0.00
0	0.01	0.02	0.02	0.07	0.12	0.29	0.17	0.18	0.05	0.07	0.00	0.01	0.00

```
##
##  Pearson's Chi-squared test
##
## data:  gp_f$sbpenage and gp_f$scityl
## X-squared = 139.15, df = 85, p-value = 0.0001931
```

The proportion table above shows the distribution of first penetrative sex age of female students in different level cities. ***The row proportion distribution actually seems pretty similar across different levels.*** And the chisq-test result is statistically significant at 0.001 level, which is to say the first penetrative sex age is correlated with the city where they attended colleges. However, we are not sure yet whether such a significant result result from the unbalance of the dataset or the potential correlation between city level and first penetrative sex age.

12	13	14	15	16	17	18	19	20	21	22	23	24	25
0	0.01	0.01	0.03	0.07	0.11	0.27	0.20	0.17	0.09	0.04	0.01	0.00	0.00
0	0.00	0.01	0.04	0.07	0.14	0.29	0.18	0.15	0.07	0.03	0.01	0.00	0.00
0	0.00	0.01	0.02	0.06	0.12	0.28	0.19	0.15	0.08	0.05	0.01	0.01	0.00
0	0.00	0.02	0.04	0.06	0.12	0.30	0.19	0.13	0.08	0.04	0.02	0.01	0.01
0	0.00	0.01	0.02	0.06	0.08	0.27	0.19	0.22	0.07	0.05	0.03	0.01	0.00
0	0.01	0.02	0.02	0.07	0.12	0.29	0.17	0.18	0.05	0.07	0.00	0.01	0.00

```
##
##  Pearson's Chi-squared test
##
## data:  gp_m$sbpenage and gp_m$scityl
## X-squared = 140.56, df = 70, p-value = 1.194e-06
```

The table above shows the distribution of first penetrative sex age of male students in different level cities. ***The row proportion distribution also seems pretty similar across different levels below 20, while above 20 there seems to be some difference across different city level.*** The result of chisq-test is also significant. But similar concern as in the female samples also exist here.

2.1.2 School Tier

	12	13	14	15	16	17	18	19	20	21	22	23	24	25
associate1	0	0.00	0.01	0.04	0.08	0.18	0.37	0.18	0.07	0.03	0.01	0.01	0.01	0
associate2	0	0.00	0.01	0.05	0.09	0.13	0.38	0.16	0.09	0.04	0.02	0.01	0.00	0
associate3	0	0.00	0.02	0.05	0.12	0.17	0.44	0.12	0.06	0.01	0.00	0.00	0.00	0
bachelor1	0	0.01	0.01	0.03	0.04	0.10	0.29	0.20	0.18	0.08	0.03	0.01	0.01	0
bachelor2	0	0.00	0.02	0.03	0.06	0.11	0.23	0.23	0.19	0.08	0.03	0.01	0.00	0
bachelor3	0	0.00	0.01	0.02	0.06	0.10	0.25	0.20	0.18	0.09	0.05	0.02	0.01	0
bachelor4	0	0.00	0.01	0.04	0.08	0.15	0.28	0.16	0.15	0.07	0.04	0.01	0.01	0

```
##
##  Pearson's Chi-squared test
##
## data:  gp_f$sbpenage and gp_f$slevel
## X-squared = 363.61, df = 102, p-value < 2.2e-16
```

Above is the distribution of first penetrative sex age of female students in different level schools. ***It seems like female students in associate's degree institutions tend to have sex at a younger age (below 19).*** The result of chisq-test is also significant. Similar concern as above tests also exist here.

	13	14	15	16	17	18	19	20	21	22	23	24	25
associate1	0.00	0.01	0.03	0.10	0.24	0.38	0.13	0.07	0.02	0.01	0.00	0.01	0
associate2	0.01	0.02	0.06	0.17	0.21	0.31	0.12	0.06	0.02	0.01	0.01	0.00	0
associate3	0.02	0.03	0.07	0.15	0.23	0.36	0.06	0.05	0.02	0.00	0.01	0.00	0
bachelor1	0.01	0.01	0.01	0.05	0.11	0.27	0.23	0.16	0.09	0.04	0.01	0.00	0
bachelor2	0.01	0.02	0.03	0.08	0.10	0.26	0.20	0.19	0.08	0.03	0.01	0.00	0
bachelor3	0.00	0.02	0.04	0.09	0.12	0.27	0.16	0.16	0.07	0.05	0.01	0.01	0
bachelor4	0.01	0.02	0.04	0.10	0.17	0.31	0.13	0.11	0.06	0.03	0.01	0.01	0

```
##
##  Pearson's Chi-squared test
##
## data:  gp_m$sbpenage and gp_m$slevel
## X-squared = 455.88, df = 84, p-value < 2.2e-16
```

Above is the distribution of first penetrative sex age of male students in different level schools. ***Same as above, male students in associate's degree institutions also tend to have sex at a younger age (below 19).*** The result of chisq-test is also significant. Similar concern as above tests exist here.