Sex, Drugs, and Rock & Roll Survey in a First-Year Service Course in Statistics

W. M. BOLSTAD, Lyn A. HUNT, and Judith L. MCWHIRTER

Sex, drugs, and rock & roll are topics that arouse almost universal interest among students. We use that interest to motivate the learning of statistical thinking in our first-year Introduction to Statistical Methods course. We have designed a class survey that uses randomized responses. Each person does his/her own randomization with a pair of colored dice. This engages the students in the gathering and analyzing of sensitive data about his/her class in such a way that no individual's personal information is divulged. The survey gives the students a concrete situation where randomization removes the motivation to give false answers to sensitive questions. Instead, the "false" answers come into the data by the randomization which has a known probability structure, so the students can extract estimates of the class proportions without knowing any individual's personal information. This reinforces the message that statistical methods can only be relied on to give meaningful results when there is some known random structure underlying the data.

KEY WORDS: Data analysis; Privacy; Randomized response; Statistical thinking.

1. INTRODUCTION

The only contact many students have with statistics comes in the first-year service course, which means this is the only chance we have to engage them in statistical thinking. Our course covers data analysis techniques, the need for randomization in the collection of data for surveys and experiments, elementary probability, and statistical inference for means and proportions. We teach this course each semester to about 150 students, and in the summer semester to about 30 students. There are three lectures each week for a 12-week semester. In addition to the lectures each student chooses one of the six practical sessions to attend every week. Most of the students in the course are in their first year, and come mainly from the computer science, science, or social science faculties.

In the practical sessions, the students get an opportunity to encounter statistical ideas in a practical setting. The students are divided into small teams. Usually, the session starts with the teams collecting their own data in a hands-on way. Sometimes they perform a small experiment, or they may select a sample from a finite population. The analysis required each week is designed to use techniques from the previous week's lectures. We encourage group discussions during these sessions, guided by the tutor. Team members are encouraged to explain their ideas to each other; however, each student writes up his/her final analysis individually, as one of our goals is to develop the students' written communication skills.

A comment frequently heard in service courses is that the course would be more meaningful for the students if they had a dataset relevant to topics that engage them. We believe sex, drugs, and rock & roll are topics never too far from the typical student's thoughts and harness this interest in the service of teaching statistical concepts.

Sometimes articles in the popular press highlight statistics on the number of sex partners a person has had or on the proportion of people who use drugs. Many people are interested in these articles, which are often headlined on the cover as an enticement to buy the publication. Frequently, the article contains no information on how the data were collected, and the statistics are stated as if they apply to the population in general. Often the data were collected using self-selected samples. Many statistical textbooks use these surveys as starting points for discussion on population, samples, and random sampling as the basis for inference. For example, Berry (1996) critiqued *Cosmos Girl*'s sex study and Wild and Seber (1999) and Kitchens (1996) discussed the Hite Report to illustrate fundamental issues in statistical thinking on a topic that interests students.

We decided that our students would definitely be interested in reliable information about their cohort and have designed a randomized response survey that we have used very successfully for the past few years. In this survey, we estimate the distribution of the number of sex partners for the population consisting of that class. Also, we estimate the marijuana usage proportions for that class. The number of sex partners a person has had and his/her previous marijuana usage are very sensitive pieces of personal information. Students clearly see that there could be a temptation to give false answers to these questions, as many people may be very reluctant to divulge their information. Our survey is done in such a way that there is no motive not to answer truthfully. If they lie, they are only lying to themselves. Since the phrase "sex and drugs" sounds incomplete without "rock & roll," we also include a question where they are asked to select the "greatest rock & roll singer of all time" from our list, just for fun. This is also answered according to randomized response even though the answer is not sensitive.

2. PRIVACY ISSUES

We had discussions with members of the university ethics committee beforehand to ensure that our protocol fully complied with university policy. We are gathering sensitive information from individuals in the class. This has to be done in a way that fully respects the students' privacy. Our protocol does so, in a number of ways.

Bill Bolstad is a Senior Lecturer, Lyn Hunt is a Lecturer, and Judith McWhirter is a Lecturer, Department of Statistics, University of Waikato, Hamilton, New Zealand (E-mail: bolstad@waikato.ac.nz). We thank the editor and referees for their constructive comments.

• The survey is anonymous as there are no names on the survey sheet. The only other information collected is the student's gender which is used for comparison purposes. The area where the survey is done is in a corner of the tutorial room, away from other people and all sheets are filled in with the same pen.

• Each student performs his/her own randomization using the dice and either answers the dummy question or the sensitive question as directed. This is explained in Section 3. The sensitive and dummy questions have the same range of answers. Thus, the answer does not divulge the personal details, since no one else knows which question was answered. The last question on rock & roll is a further privacy safeguard to ensure that the dice are not inadvertently left in such a way that shows whether the sensitive question on marijuana usage was answered.

• The person folds his/her own answer sheet and puts it in a sealed "hand-in" box. After all sheets are collected, the Departmental Secretary opens the box, collates the results, and destroys the original sheets.

We also indicate to students that participation is purely voluntary. If they do not wish be singled out for nonparticipation, they can participate and put x instead of a number. We do emphasize however, that they must decide before they roll the dice whether or not they are going to answer the sensitive question if the randomization requires it. They should not wait until the dice indicates they are to answer the sensitive question and then decide to put an x.

3. THE SURVEY PROTOCOL

The survey data collection is carried out in a practical session, while another problem is being worked on by the class. One corner of the room is set up for the survey, with a table and sealed "hand-in" box. Two six-sided dice, one red and one green, are in the cup on the table. Each participating student goes to the area when the previous student has finished. It only takes each student one to two minutes to do the survey, and his/her participation does not interfere with other work going on during that session.

Each student performs three randomizations, one for each question. First he/she shakes and rolls the two dice, keeping them covered by the cup. The cup is then tilted so only he/she can see the numbers rolled. When the student finishes the sex partner question, he/she rolls the pair of dice for the marijuana usage question. After completing that question, he/she rolls the dice a third time, for the rock & roll question. For each roll, the number on the red die determines the course of action taken by the student, for that question.

Roll One

If the red die shows a 1 or 2, the answer is determined by the number on the green die.

If the green die shows 1	Put 1
If the green die shows 2	Put 2
If the green die shows 3	Put 3
If the green die shows 4	Put 4
If the green die shows 5	Put 5
If the green die shows 6	Put 0

or else

If the red die shows 3, 4, 5, or 6, answer the following question:

The number of sex partners I have had is

0	Put 0
1	Put 1
2	Put 2
3	Put 3
4	Put 4
5 or more	Put 5

My Answer is _____

Roll Two

If the red die shows a 1 or 2, the answer is determined by the number on the green die.

If the green die shows 1, 2, or 3	Put 0
If the green die shows 4 or 5	Put 1
If the green die shows 6	Put 2

or else

If the red die shows 3, 4, 5, or 6, answer the following question:

The most recent time I have tried smoking marijuana is

Never	Put 0
Before this semester	Put 1
During this semester	Put 2

My Answer is _____

Roll Three

If the red die shows a 1 or 2, the answer is determined by the number on the green die.

If the green die shows 1	Put 1
If the green die shows 2	Put 2
If the green die shows 3,	Put 3
If the green die shows 4	Put 4
If the green die shows 5	Put 5
If the green die shows 6	Put 0

or else

If the red die shows 3, 4, 5, or 6, answer the following question:

The greatest Rock & Roll singer of all time was

Kurt Cobain	Put 1
Jimi Hendrix	Put 2
John Lennon	Put 3
Freddy Mercury	Put 4
Jim Morrison	Put 5
Elvis Presley	Put 0

My Answer is _____

Return the dice to the cup.

After each randomization, the student writes his/her answer to the question in the appropriate box on the sheet. When all three questions have been completed, he/she folds the survey sheet and places it in the sealed "hand-in" box.

4. ANALYZING THE DATA

The results of the survey are analyzed during the practical sessions the following week. The students are guided through the use of statistical methods to find the answers to the questions posed about their class.

• For instance, do we consider the class as a single population of students and analyze the results all together, or should we consider the class as two subpopulations—male students and female students—and analyze these populations separately? Students can understand that the course we should take depends on whether males and females have different patterns in this aspect of their behavior. If they are essentially the same, the combined population gives rise to a larger amount of data, and thus will give more reliable results. However, if males and females have fundamentally different patterns, analyzing them together would obscure the differences between the subpopulations. This leads to the next question.

• Are the *number of sex partners* distributions different for males and females? If the proportion showing i sex partners is the same for males and females for all $i = 0, \ldots, 5$, then the classifications are independent. This is used to motivate the Pearson's chi-square test for independence. We use this to decide whether the benefits of pooling are worth pursuing. Because of the randomization, approximately one third of the answers are to the dummy question. With a class of 150 students, this means approximately 100 are answering the sensitive question. Dividing this in half to analyze males and females separately gives quite a small amount of data for each gender, which leads to less reliable estimates. In our classes, the chi-square test has always ended up recommending using the pooled results, with a p value that is greater than .50. Even though the power of the test has been reduced by the randomization in the data, we believe this is a strong indication that there is only a very small difference, if any, between the male and female results. Nevertheless, many students also calculate the separate estimates for males and females out of their own interest.

• How can estimates of the population proportion having i sex partners for $i = 0, \ldots, 5$ be obtained, when the data contain incorrect answers (errors)? Some of the answers are to the dummy question, but we have no idea which ones. This motivates discussion on the fundamental basis of statistical inference. It is based on knowing the probabilistic nature of the errors. When we know that, statistical methods can be developed that compensate for the errors, and in this case we do know them, since they are based on rolls of fair dice. We give two explanations that give the same estimates for the population proportions.

First, using a heuristic argument, for each i, we calculate the number of people that would be *expected* to give answer i from answering the dummy question. This *expected* number is then subtracted from the number that did give answer i to give a_i, the *adjusted* number for answer i. That is, the adjusted number for i sex partners is given by

$$a_i = n_i - n \times \frac{1}{3} \times \frac{1}{6},$$

where n_i is the number that gave answer i, n is the total number participating in the survey, $1/3$ is the proportion expected to answer the dummy question, and $1/6$ is the proportion expected to give answer i if they answer the dummy question. The adjusted number a_i is an estimate of the number who answered i to the sensitive question. The sum of the adjusted numbers is given by

$$\sum_{i=0}^{5} a_i = \sum_{i=0}^{5} n_i - 6 \times n \times \frac{1}{3} \times \frac{1}{6} = \frac{2}{3} \times n.$$

The estimate of each population proportion having i partners is the adjusted number for i partners divided by the sum of the adjusted numbers. This can be rearranged to give the formula

$$\hat{\pi}_i = \frac{a_i}{\sum a_i} = \frac{3}{2}\left(\frac{n_i}{n} - \frac{1}{3} \times \frac{1}{6}\right), \tag{1}$$

where n_i/n is the proportion of people giving answer i, $1/3 \times 1/6$ is the probability of answering i due to the dummy question, and $3/2$ is the reciprocal of the probability of answering the sensitive question.

Second, using a more formal probabilistic approach, we let θ_i be the probability a randomly selected person in the class gives answer i. Because of his/her randomization this is

$$\theta_i = \frac{2}{3}\pi_i + \frac{1}{3} \times \frac{1}{6}, \tag{2}$$

where $2/3$ is the probability of answering the sensitive question, π_i is the proportion of the population who have had i sex partners, $1/3$ is the probability of answering the dummy question, and $1/6$ is the probability of giving answer i to the dummy question. So the number of people answering i has expected value $n \times \theta_i$. Substituting back in, and rearranging to give an estimate for π_i also gives us Equation (1). At the first-year level we do not attempt to quantify the standard error associated with the estimates, although the students have met this concept and understand that increasing the number of observations decreases the standard error.

• What should we do if one of the estimates is negative? It is possible that, due to chance, an adjusted number is negative, which would give a negative estimated proportion. This leads to discussion on the common sense approach of giving up unbiasedness if it leads to a negative probability estimate, which is out of bounds.

We repeat the same procedure for marijuana usage and for choice of rock & roll singer. Because the randomization for the choice of rock & roll singer is similar to that used for the sex-partner question, we can use equation (1) to obtain the estimates for the population proportion of selecting *singer i* as the "greatest rock & roll singer of all time."

The randomization used for the marijuana usage question is more complicated. When we first instituted this survey, we asked the question *Have you ever tried marijuana* and there were only two possible answers, *yes* (1) and *no* (0). We assigned each of these responses the probability of $3/6$ in the dummy question. We decided it would be more interesting to try to partition the proportion who had tried marijuana into current and past users. We defined a current user as one whose most recent usage was during this semester, and a past user as one whose most recent usage was previous to this semester. Both *before this semester* and *during this semester* correspond to *yes* from the previous formulation of the question. To allow backward comparability and since the die has six faces, we allocated 1 to *before this*

semester and 2 to *during this semester* with associated probabilities of $2/6$ and $1/6$, respectively, for the dummy question. This method gives rise to a different formula for the probability a randomly selected person answers j, and for the estimate of the population proportion for each answer. For instance, for $j = 0$ (*never*) the formula for the probability a randomly selected person answers j is given by

$$\theta_j = \frac{2}{3} \times \pi_j + \frac{1}{3} \times \frac{3}{6}, \tag{3}$$

where $2/3$ is the probability of answering the marijuana question, π_j is the proportion who have never used marijuana, $1/3$ is the probability of answering the dummy question, and $3/6$ is the probability of giving answer $j = 0$ to the dummy question. Similarly, the estimated proportion for $j = 0$ (*never*) is given by

$$\hat{\pi}_j = \frac{3}{2}\left(\frac{n_j}{n} - \frac{1}{3} \times \frac{3}{6}\right), \tag{4}$$

where n_j/n is the proportion of people giving answer j, $1/3$ is the probability of answering the dummy question, $3/6$ is the probability of answering 0 to the dummy question, and $3/2$ is the reciprocal of the probability of answering the sensitive question. The corresponding formula for $j = 1$ (*before this semester*) and $j = 2$ (*during this semester*) are found by setting the probability of answering j to the dummy question to be $2/6$ and $1/6$, respectively. It should be noted that for the question formulated in the original version, allowing answers of *yes* and *no*, Equation (4) gives the required estimates for the population proportions of both responses.

5. REVISITING THE DATA IN A SUBSEQUENT COURSE

We return to the dataset in the *Statistical Data Analysis* course that some of the students take in the following year. They have done the survey themselves, so they are familiar with the context of the data. Here we give them the combined dataset for the three years we have been doing the survey. They use Minitab to redo the analysis on the combined dataset for all the previous years, and also do some further analysis to answer some additional questions.

• How do we calculate the standard error of the estimated proportions? The estimated proportions for answer i is a linear function of n_i, the number of people giving answer i. For instance, this is given by Equation (1) for the sex-partner question. Ignoring the finite population correction factor, the variance of n_i equals

$$\text{var}(n_i) \;=\; n \times \theta_i \times (1 - \theta_i),$$

where θ_i, the probability a randomly selected person gives answer i, is given in Equation (2). This simplifies to

$$\text{var}(n_i) = n \times \left(\frac{2}{3}\pi_i + \frac{1}{3} \times \frac{1}{6}\right) \times \left(1 - \left(\frac{2}{3}\pi_i + \frac{1}{3} \times \frac{1}{6}\right)\right).$$

We use the formula for finding the standard deviation of a linear function of a random variable. The estimated standard error of the estimate is thus

$$\sigma_{\hat{\pi}_i} = \frac{3}{2}\sqrt{\frac{1}{n}\left(\frac{2}{3}\hat{\pi}_i + \frac{1}{3} \times \frac{1}{6}\right) \times \left(1 - \left(\frac{2}{3}\hat{\pi}_i + \frac{1}{3} \times \frac{1}{6}\right)\right)}.$$

Similar expressions can be found for the standard errors of the estimates for the marijuana usage estimates. For instance for $j = 0$ we would use Equations (3) and (4), and would modify them as explained in Section 4 for $j = 1$ and $j = 2$. The rock & roll question has the same randomization as the sex-partner question, so the standard errors would be given by Equation (5).

• Can we conclude that there is a difference between the true proportions of males and females reporting i partners, $i = 0, \ldots, 5$ for some particular i? This motivates the use of the two sample test for equality of proportions. The test statistic z equals the difference between male and female proportions divided by the square root of the sum of squares of the respective standard errors.

• Do the results of these tests conict with the Pearson's chi-square test done previously? This leads to discussion on multiple hypothesis testing and overall level of significance. We do the similar analyses for marijuana usage, and for "greatest rock & roll singer of all time."

• Are marijuana usage and number of sex partners correlated? Both multiple sex partners and marijuana usage have potential health risks. Is the willingness to take on one of the risks related to the willingness to take on the other? We combine the answers *before this semester* and *during this semester* finishing the current marijuana question *The most recent time I have tried smoking marijana is* to give the answer *yes* to our original marijuana question *Have you ever tried marijuana* as explained in Section 3. This allows us to use all the data we have collected, from both the current and original versions in the form of the original version variables.

We estimate the correlation coefficient the following way. We cross tabulate the answers and find the adjusted numbers by subtracting off the number that would be expected to give each answer from the randomization. The adjusted number of respondents with i sex partners and j marijuana usage number is given by

$$a_{i,j} = n_{i,j} - n * \left(\frac{1}{3} \times \frac{1}{6}\right) * \left(\frac{1}{3} \times \frac{3}{6}\right).$$

Here $n_{i,j}$ is the number answering i and j to the sex partner and *Have you ever tried marijuana* question respectively, n is the total number, $1/3 \times 1/6$ is the probability of answering i to the sex-partner question due to the randomization, and $1/3 \times 3/6$ is the probability of answering j due to the randomization to the original marijuana question . The possible values are $j = 0, 1$. The adjusted number $a_{i,j}$ estimates the number of respondents who answered i to the sensitive sex-partner question and j to the sensitive (original) marijuana question. We estimate the mean

number of sex partners by

$$\bar{x} = \frac{\sum_{i=0}^{5}\sum_{j=0}^{1} i \times a_{i,j}}{\sum_{i=0}^{5}\sum_{j=0}^{1} a_{i,j}},$$

and the mean marijuana number (proportion who have tried marijuana) by

$$\bar{y} = \frac{\sum_{i=0}^{5}\sum_{j=0}^{1} j \times a_{i,j}}{\sum_{i=0}^{5}\sum_{j=0}^{1} a_{i,j}}.$$

Similarly, we estimate the variance of number of sex partners by

$$S_x^2 = \frac{\sum_{i=0}^{5}\sum_{j=0}^{1} (i-\bar{x})^2 \times a_{i,j}}{\sum_{i=0}^{5}\sum_{j=0}^{1} a_{i,j}},$$

the variance of the marijuana usage number by

$$S_y^2 = \frac{\sum_{i=0}^{5}\sum_{j=0}^{1} (j-\bar{y})^2 \times a_{i,j}}{\sum_{i=0}^{5}\sum_{j=0}^{1} a_{i,j}},$$

and the covariance of number of sex partners and marijuana usage number by

$$S_{xy} = \frac{\sum_{i=0}^{5}\sum_{j=0}^{1} (i-\bar{x}) \times (j-\bar{y}) \times a_{i,j}}{\sum_{i=0}^{5}\sum_{j=0}^{1} a_{i,j}}.$$

Thus, the correlation is estimated by

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_x^2 \times S_y^2}}.$$

We have found that the estimated correlation coefficient between these two variables (although moderate sized) is highly significant given our total sample size. This may indicate that both variables are linked to a willingness to indulge in risky behaviors, although other possible explanations for the correlation may also exist.

• Finally we discuss what larger population the results may be applicable to. This brings in the very fundamental statistical idea that inference depends on random sampling from a population. In this case while the results may be indicative of some larger population there is no basis for any wider statistical conclusions than to the class in question. In terms of any wider population, it is an observational study since the sample (class) is self-selected by choosing one of our Introduction to Statistics courses.

6. SUMMARY

In practice we have found that the sex, drugs, and rock & roll survey is a better teaching tool than we ever imagined. We have found that the motivation of the survey can be used in many topics (observational studies versus randomized experiments, target population, etc.), and we also had a dataset the students were extremely interested in.

The idea of the survey is introduced early in the course. This gives the students time to think about participation in the survey and also stimulates interest. We show how bias can be introduced into datasets through incorrect answers, and through using self selected samples. This leads to a discussion on how we could collect a reliable dataset that we could use that would also respect their privacy. This addresses any apprehension they may initially have. By the time we are ready to conduct the survey, they are fully satisfied that their privacy is being respected and are very interested in finding out what the class results will be.

We find that almost everyone participates fully in the survey. They realize that their information is necessary to get the full picture of the class, and willingly contribute. The few *x*'s that occur are almost always on the rock & roll question where they think our selection does not contain their favorite!

Some students are quite emphatic that they would not mind giving their personal information, regardless of the randomization. However, they can see the importance of following the protocol when we point out the similarities to an experiment where the treatment group gets a new drug that may reduce tumor size and the control group gets a placebo. All patients in the study would hope to be assigned the new drug. However, its effectiveness could never be determined unless the randomization protocol is strictly followed. The students gain experience in following the protocol of a randomized experiment and see how statistical techniques can be used to collect sensitive information. They also see how we can extract the estimates of the parameters of the population when "false" answers are going in with known probabilities.

We think that 100 is about the minimum class size needed to get meaningful results. For any class below this size, we would recommend combining with other class results. In our summer semester the class size is below this threshold, so we combine the results with those from the previous semester. In conclusion, the sex, drugs, and rock & roll survey is a very useful teaching tool in an Introduction to Statistics Class.

[Received November 1999. Revised May 2000.]

REFERENCES

Berry, D. A. (1996), *Statistics: A Bayesian Perspective*, Pacific Grove, CA: Duxbury.

Kitchens, L. J. (1996), *Exploring Statistics: A Modern Introduction to Data Analysis and Inference*, Pacific Grove, CA: Duxbury.

Wild, C. J., and Seber, G. A. F. (1999), *Chance Encounters: A First Course in Data Analysis and Inference*, New York: Wiley.