How to ask sensitive questions using statistics: a case study on academic dishonesty

Gina Londino and Connie Waung

Gina Londino graduated from Ball State in May of 2004 with a major in Chemistry and a minor in Mathematics. Gina will be continuing her education by obtaining a Masters' of Chemistry at Indiana University-Purdue University in Indianapolis in the fall. Before then, Gina will be interning at Eli Lilly and Company for the summer.

Connie Waung is a graduate student at Ball State obtaining her license in secondary education with a major in Business (education) and minor in Mathematics (education). She currently holds a BSBA from The Ohio State University as well as an MBA from Wayne State University. She has 12 years experience working for DaimlerChrysler in business management.

Our project focus was to determine the proportion of students who have cheated on a test at least once in the past year. Out of the students that cheated we were then to determine if a student was more likely to be a freshman, sophomore, junior or senior.

The first thing we had to do, was to define *cheating*. The following is the definition of cheating, which we used for our project: you have cheated if you (1) copied answers from someone else on a test; (2) turned in a paper that you did not write; (3) used an unauthorized "cheat sheet"; or (4) discussed the answers on take home test with somebody else [3].

Once we had decided on what was meant by cheating, we had to determine how to collect the data for our project. Since cheating is a "sensitive" subject, we were afraid some students would be unwilling to respond truthfully. Therefore, we had to use a type of survey that would respect students' privacy. The survey technique which we chose, allowing students to answer truthfully without having their privacy invaded, was that of the *Randomized Response Survey* (RRS).

The RRS is used in surveys dealing with sensitive questions such as drug use, cheating, etc. Many potential respondents refuse to answer, or give false information. With the RRS methodology, their responses are guaranteed to be private [1].

The procedure for RRS is as follows: (1) the respondent is faced with a sensitive question; (2) the respondent is then given some randomization device such as a coin; (3) the respondent flips the coin without showing it to the interviewer. If the coin lands heads, then the respondent answers "yes" to the question. If the coin lands tails, then the respondent answers truthfully to the question; and, after all data has been collected, (4) one computes the desired probability of a person's *correct* response being "yes", using a certain probability formula [2].

For our experiment we had only one sensitive question to ask the students and that was if they had ever cheated on a test. We first asked the students their class rank (i.e. are they a freshman, sophomore, junior or senior?). Next, we gave the students a paper and explained to them how the survey would work and gave them our definition of cheating. We then gave the students a coin and told them to flip the coin and answer "yes" to our question if the coin was heads and answer truthfully to the question if the flip of the coin was tails. The students were instructed not to show us the outcome of the coin flip and we explained to them that this would respect their privacy (i.e. we would not know if they said "yes" because they had flipped heads or if they said "yes" because they had cheated).

After the students flipped the coin, we reminded them "yes" for heads, "truth" for tails, and asked them the question "Have you cheated on a test in the past year (October 2002-October 2003)?" After the students answered, their answer was recorded and they were offered candy for their participation. The total number surveyed was 38 students. Here are the results:

	Freshmen	Sophomores	Juniors	Seniors
Total Surveyed	9	9	9	11
Yes Reponse	7	5	5	7

Once we had completed surveying students, we calculated the overall probability of cheating, using the following formula, taken from [4]:

$$P(\text{yes}) = P(\text{yes} \mid \text{heads})P(\text{heads}) + P(\text{yes} \mid \text{tails})P(\text{tails}).$$

Here, $P(\text{yes})$ is the probability that a student would answer "yes" to our question. This probability can be estimated by the proportion of students who responded "yes" to the survey. $P(\text{yes} \mid \text{heads})$ is the conditional probability that a student says "yes" when the flip is heads. This probability equals 1, because a student automatically responds with "yes" in this case. $P(\text{yes} \mid \text{tails})$ is the conditional probability of saying "yes" to tails, which is the number we are trying to estimate: the probability that a student actually has cheated. $P(\text{heads})$ and $P(\text{tails})$ are the probabilities of obtaining heads and tails, respectively, and they are both equal to $\frac{1}{2}$. If we let $P(\text{yes} \mid \text{tails}) = P$ in the above formula, and solve it for P we get:

$$P = 2P(\text{yes}) - 1. \tag{1}$$

In our survey, 24 students from a total of 38 gave a "yes" response, so we estimated the probability of cheating to be

$$P = 2\left(\frac{24}{38}\right) - 1 = \frac{5}{19}.$$

Therefore our results suggest that the proportion of students who have cheated on a test at least once in the past year is approximately 26%.

After we determined the overall cheating probability, we wanted to determine whether a student who has cheated is more likely to be a freshman, sophomore, junior or senior. The following is our null hypothesis.

> **Null Hypothesis**: If a student has cheated on a test, that student is equally likely to be a freshman, sophomore, junior or senior (25% freshman, 25% sophomore, 25% junior, 25% senior).

We will test this hypothesis using the chi-square test [5]. Below are the results and calculations we need to apply the test.

> **Freshman Results**: The conditional probability that a student is a freshman, given that the student has cheated, is given by: $P(\text{freshman} \mid \text{cheat}) = P(\text{cheat} \mid \text{freshman}) \cdot P(\text{freshman})/P(\text{cheat})$, another formula which we learned from [4].

$P(\text{cheat} \mid \text{freshman})$ can be computed from Formula (1), by only considering the results of the survey as it relates to the freshman in computing $P(\text{yes})$. For example, in our survey there were 9 freshman and 7 of them responded yes. So for the freshman sample only, $P(\text{yes}) = \frac{7}{9}$. If we use this probability in (1), we obtain $P(\text{cheat} \mid \text{freshman}) = \frac{5}{9}$. We also note that $P(\text{freshman}) = \frac{9}{38}$ and $P(\text{cheat}) = \frac{5}{19}$, so that:

$$P(\text{freshman} \mid \text{cheat}) = (\frac{5}{9} \cdot \frac{9}{38})/(\frac{5}{19}) = \frac{1}{2}.$$

In other words, if a student has cheated, he/she is a freshman with a likelihood of 50%.

We then followed the same logic to calculate the conditional probabilities for sophomores, juniors, and seniors. Our results are summarized in the following table.

Observed distribution of cheaters:

Freshmen	Sophomores	Juniors	Seniors
50%	10%	10%	30%

To see whether the apparent differences seen in the table above are statistically significant or not, we computed the chi-square statistics [5]. Note that the "O" and "E" values in the table below correspond to observed and expected frequencies. We estimate the total number of cheaters in our study to be $P(\text{cheat}) \cdot 38 = \left(\frac{5}{19}\right) \cdot 38 = 10$. Here are our results:

Rank	O	E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
Freshmen	$50\% \cdot 10 = 5$	$25\% \cdot 10 = 2.5$	6.25	2.5
Sophomores	$10\% \cdot 10 = 1$	$25\% \cdot 10 = 2.5$	2.25	0.9
Juniors	$10\% \cdot 10 = 1$	$25\% \cdot 10 = 2.5$	2.25	0.9
Seniors	$30\% \cdot 10 = 3$	$25\% \cdot 10 = 2.5$	0.25	0.1
Total	10	10		4.4

Chi Square Value = 4.4

The probability that the chi-square statistic with 3 degrees of freedom gives a value larger than 4.4 is more than 0.05. Therefore, we cannot reject the null hypothesis at a 5% significance level, and accept that if a student has cheated on a test, then the student is equally likely to be a freshman, sophomore, junior or senior.

This project was part of a term paper in MATHS 222. The faculty advisor for the project was Dr. Giray Ökten.

References

[1] W. Du, and Z. Zhan, *Using Randomized Response Techniques for Privacy-Preserving Data Mining,* Department of Electrical Engineering and Computer Science of Syracuse University. `http://www.sai.syr.edu/facultypapers/Randomized%20Response%20Techniques.pdf`

[2] V. Lesser, *Asking Sensitive Questions and Questionnaire Format*, Notes of a lecture given on March 10, 2003, and posted on the internet.

[3] Ball State University Newscenter, *Surveys find that half of college students admit to cheating*, (October 28, 1998). `http://www.bsu.edu/news/article/0,1370,-1019-635,00.html`

[4] G. Ökten, Conditional Probability and Survey Designs: Randomized Response Technique, Lecture Notes for Maths 335, Ball State University, pp. 67–72.

[5] D. Monrad, W.F. Stout, E.J. Harner, B.A. Bailey, R.L. Gould, X. He, L.A. Roussos, J.S. Colburn, Statistics: the craft of data collection, description, and inference (Third Edition), Mobius Communications (2002).