STATISTICS PROJECTS USING INSTITUTIONAL DATA

THOMAS J. PFAFF

Abstract. If we let all students at Ithaca College be our population, then our Office of Institutional Research can provide us with various parameters about this population. For example, we can obtain parameters regarding SAT scores, birth month, and GPA. Each student samples from the population and we compare their results to the parameters. This allows us to better illustrate the meaning of confidence intervals, P-values, sample Size, the power of a test, and the central limit theorem. We provide examples of what was done in class using the data the students collected.

1. Introduction

At Ithaca College, as at most other schools, there is an Office of Institutional Research that collects and maintains a wide range of data. Thus, if the student body is viewed as our population, we have parameters related to SAT scores, GPA, birth month, and more. This provides us with the opportunity to have students sample attributes of the student population and compare their results to the actual parameters. In this paper, we provide examples of using institutional data to illustrate the meaning of confidence intervals, P-values, sample size and power, and the Central Limit Theorem. Along with deepening students' understanding of these topics, each student also gained experience in collecting data from human subjects. Separately, students also did a group project in the course, but not everyone chose a project that has them sampling people. The projects here ensures that each student will gain experience sampling from human populations.

The student population at Ithaca College is approximately 6000 students. The data the class collected was gender, birth month, SAT math scores, SAT verbal scores, and the previous semesters GPA. The parameters associated with birth month and SAT scores were requested from our Office of Institutional Research. The parameter for GPA data was already

Key words and phrases. Statistics, sample, confidence interval, power.

available to faculty on our registrar's web page. The rest of this paper contains an overview of what was done in our project or could be potentially done. Everything that follows is something that was discussed in class except the commentary on power and the Central Limit Theorem, which will be included in the future.

2. Data Collection

Before asking students to collect data, we had to obtain approval from our human subjects review board. In getting permission to collect this information we had to insure our review board that all information would be kept confidential. When collecting the data, students were informed to not record the telephone number with the collected data. In fact, this survey is essentially anonymous since you will notice in the protocol below that students never actual ask for the name of the student they contact. They could be speaking to a roommate or friend as opposed to the person on their list. Finally, you will notice in the protocol that it is made clear that one can opt out of any and all questions at any point in the interview.

Once the approval was secured, students were expected to obtain sample information of gender, birth month, SAT math scores, SAT verbal scores, and the previous semesters GPA from twenty students. A sample of size twenty is large enough to be useful and small enough to be obtainable. In practice, students wouldn't always get a full sample of size twenty, since some students would occasionally opt out of a question. First, students generated a random number from 2 to 5948, which corresponded to a list of registered students from the registrar's office. It was necessary to obtain this list from the registrar's office specifically for this purpose, since the college no longer prints a hard copy campus directory. The list contained each student's name, local phone number, and email address. Once they had selected numbers from this list, students used the following protocol when making phonecalls for obtaining their sample data:

> Hi, I'm a student in business statistics and we are conducting a survey. I would like to ask you four questions and you may refuse to answer any or all of the questions. Your answers will be kept confidential, as they will not be

linked to the phone number that I have called. Moreover, I will not ask for your name in this survey. Are you willing to participate?

(Yes) Thank you. First, are you 18 years or older and an Ithaca College student?

(Yes) Again let me remind you that you may refuse to answer any of these questions.

1) Are you male of female?

2) What month were you born in?

3) What are your math and verbal SAT scores?

4) What was your GPA last semester?

Thank you for participating.

Students were instructed to try contacting an individual up to 6 times, varying the time and day called, before giving up. Students were also advised to try and contact the person via email if they failed by phone. If the phone number in the directory was incorrect students checked the online directory, which is more up to date, before moving to another number.

Overall, this process went smoothly, and there was often discussion before class about the trials and tribulations of obtaining the sample. In particular, students gained an appreciation of the issue of how individuals refusing to respond may effect the sample. One semester, at about the same time we were collecting the data, one of the students was being called by national polling organization. Despite repeated attempts the student wouldn't respond. Interestingly, it seems going through this process doesn't necessarily make one sympathetic to pollsters.

3. Confidence Intervals

Confidence intervals were constructed for the proportion of students born in January through May, mean SAT math scores, and mean SAT verbal scores. The purpose here was to help the students understand the meaning of a confidence interval, beyond a picture in a book (see Figure 6.2 in [1]) or an applet, and to see what can go wrong.

The first confidence interval we looked at was for the proportion of students born in January through May, because these intervals are well behaved, students know what month they were born in and provide that information accurately, as compared to SAT scores. We chose January through May for no particular reason other than to avoid a proportion of 0.5, since students might think there is something special about 0.5. The proportion of Ithaca College students born in these months is 41.95%. Students were asked to construct an 80% confidence interval for the percentage of students in their sample that were born in January through May, and the intervals were written on board. The creation of the confidence interval was done by the students on a TI-83 calculator. The expectation here is that each student has an 80% chance of containing the true proportion in the interval. We then did a 95% confidence interval and again put these on the board. A summary of the results of these z-intervals are given in Table 1.

Summary CI Proportion			
	% Correct	% Over	% Under
80% CI	86.2%	6.9%	6.9%
95% CI	96.5%	0.0%	3.5%

Table 1. Summary of the 80% and 95% confidence intervals for the proportion of students born in January through May.

In our discussion of the results we noted that for an 80% confidence interval we expected about 20% of the class to miss the true proportion, and that we should have about an equal balance between intervals that were above or below the true proportion. We had a similar discussion about the 95% intervals, along with noting the differences between the two types of intervals. In particular, we noted that the 95% intervals were wider.

The next intervals were 80% and 90% intervals for the mean SAT math score. Here we used a standard t-interval, and the campus mean is a score of 588. A summary of the results of these intervals are given in Table 2. In this case, the results are not as expected and, in fact, more interesting.

Summary CI Mean SAT Math			
	% Correct	% Over	% Under
80% CI	51.7%	41.4%	6.9%
90% CI	65.5%	27.6%	6.9%

Table 2. Summary of the 80% and 90% confidence intervals for the mean SAT math score.

Our discussion of the SAT math intervals started with trying to explain why we missed the parameter much more than expected. The students in the class noted quickly that the people they were interviewing typically paused, sometimes at length, before answering the question on SAT scores. They did not seem to notice right away the disproportionate number of intervals that overestimated the actual mean. For the 80% confidence interval, 41.4% of the intervals missed the true mean by overestimating the mean, while only 6.9% underestimated the true mean. In this case we had 14 intervals that missed the true mean, with 12 over and 2 under, providing an opportune time to review binomial probabilities. Assuming that there should be an even split, the probability of having 12 or more out of 14 over is 0.0065. In other words, it is highly unlikely this happened by chance.

A conversation on the honesty of our sample results ensued. Did our subjects consciously lie to us about their SAT scores? Or, was this simply due to faulty memories? Could there be other explanations? To provide an alternative explanation, we discussed that SAT scores on campus have been increasing and that we may have more first year students in our sample than we might expect. This could explain some of the results. In the future, class year will be included in our list of sample questions. Our results for the SAT verbal mean were similar (see Table 3) and expected once we did the SAT math scores. Overall, these discussion were both engaging and instructive for the students.

4. P-values, Sample Size, and Power

Here we begin by having the students calculate the P-value for a 2-sample t-test comparing female GPA against male GPA. Students would then write their P-values and whether males of females had the larger sample mean. This gives them an opportunity to see how the

Summary CI Mean SAT Verbal			
	% Correct	% Over	% Under
80% CI	37.9%	51.7%	10.3%
90% CI	55.2%	37.9%	6.9%

Table 3. Summary of the 80% and 90% confidence intervals for the mean SAT math score.

randomness of the sample affects the P-values (see Table 4). Although there were only six P-values below 0.05, the fact that 75% of our female sample means were larger than the corresponding male sample means suggests that the females have higher GPA's. This is, in fact, true with the females outscoring the males by 0.25 on our campus.

P-values testing Male GPA vs Female GPA					
0.012	0.089	0.681	0.047	0.330	0.710
0.307	0.520	0.046	0.668	0.038	0.788
0.083	0.170	0.891	0.731	0.380	0.258
0.652	0.018	0.151	0.445	0.306	0.907
0.882	0.960	0.166	0.786	0.042	

Table 4. P-values testing Male GPA vs Female GPA, with a parameter of 0.25 = FGPA - MGPA. At $\alpha = 0.05$ the rejection rate is 21% compared to a power calculation of 25% ($n = 9$, $s = 0.39$).

This is an opportune time to discuss the power of a test. Why did only 21% of the P-values detect a difference? Using Minitab one can easily find the power of our test. Estimating a standard deviation from our sample, we get powers of 22%, 25%, and 27% for sample sizes in each group of 8, 9, and 10 respectively. Our groups didn't have equal sample sizes, and even though students attempted to collect a sample of size 20 we had some missing GPA's. Hence, it is fair to look at sample sizes of less than 10 for each group. One can also use the power calculator by putting in the sample size, rejection rate, and standard deviation to obtain a difference. Using $n = 9$, power$= 0.21$, and $s = 0.39$, we get a difference of 0.22.

Based on either calculation, our success rate at detecting a difference was what we should expect.

Whether one talks about power or not, it is instructive to then group the students together and pool their data. The key idea is that the larger the sample size the more likely we are to detect difference and reject the null hypothesis, when it is false as we know it is. Note that by pooling our data this way we could have the same person sampled twice. With our data, if we make groups of size three (leaving two data sets out for our discussion here) we reject equal GPA's in four out of the nine cases, with one of the P-values at 0.09. Again this is close to what is predicted by the power of the test. An $n = 25$ gives a power of 60%. We rejected 44% of the time (55% with one close call, 0.09). Also, all nine of the female sample means were larger. If we triple the groups again to three groups of size nine we just miss all three groups rejecting with the largest p-vlaue of 0.0550, and the power close is to 100%.

5. The Central Limit Theorem

We can also use our data to illustrate the Central Limit Theorem. Using the SAT math data, if we average the 29 standard deviations from each student we get 67.6. The actual standard deviation of SAT math scores is 68.39. We can now calculate the standard deviation of the 29 sample means, which is 18.71. According to the Central Limit Theorem we should get $68.39/\sqrt{20} = 15.29$. This is despite the fact that there seems to be some inflation in our SAT math scores. One might conclude that students are consistent in their inflation of scores.

We can do a similar analysis with the SAT verbal scores. The average of the 29 standard deviations from each student is 70.63, with an actual standard deviation of SAT verbal at 73.36. We can now calculate the standard deviation of the 29 sample means, which is 25.43. According to the Central Limit Theorem we should get $73.36/\sqrt{20} = 16.40$. Again our results are close to what is predicted.

6. Conclusion

Overall, these class projects seem to help students understand the randomness behind the statistics. Once the students collect the data, which can be done early in the course, these projects and discussion can be integrated whenever appropriate. It is also helpful to collect the data from the students by having them send it in some standard electronic format, such as an Excel file. There is a positive impact on the student's learning from the fact that the data is real and clearly not fabricated, since they collected the data themselves. Of course, this keeps the instructor on edge since we are dealing with live data so one has to be prepared to deal with the unexpected. Thankfully though, randomness tends to behave as expected.

Acknowledgement Thanks to Claire Powers in our Office of Institutional Research who kindly provides me with various parameters every spring.

References

[1] Moore, D., McCabe, G., Duckworth, W., & Sclove, S., *The Practice of Business Statistics*, W. H. Freeman, New York, 2003.

Tom Pfaff is an assistant professor of mathematics at Ithaca College. When not teaching statistics courses he is busy, along with his wife, raising their four sons. This sample of size four, is plenty large enough to be interesting, but not large enough draw any conclusions about child development. He stays in shape by rowing, running, cycling, (in fact, he has a B.S. in exercise science) and yes chasing kids. More about him can be found at http://www.ithaca.edu/tpfaff.