Project Part 1

Objectives and Background

Past UBC grade distributions are available online. On this public GitHub repo, course grade data are sorted into folders by year (from 1996 to 2021) and session (summer or winter). Each folder contains one table for each department, each of which contains the class size, class average, and the number of students who received a grade from 90-100, 80-90, 70-80 (etc) for every course offered in that department in that session. (More information about the dataset's structure can be found in the Appendix.)

These public grade distributions have long been used by students as a resource for making course decisions. Students are eager to know past course averages, as this allows them to make informed decisions regarding their timetables and instructors. They're also eager to see how their grades compare to their peers. In recent years, there has been some controversy regarding grade inflation at schools across North America; the UBC grades dataset allows for comparison of past and present grades. In light of recent grade inflation, is an A (80%) still a rare grade? For this project, our objective is to determine (1) the mean grade and (2) the proportion of students with a final grade of above 90% across all courses during the most recently available (at the time) 2021 winter session. This is important because knowing this will allow students to compare their grades to the global averages. We chose the mean grade because it is the most common population grade metric (e.g., GPA uses the mean; course averages are listed on students' transcripts), and we chose the proportion above 90% because this is something that most motivated students are interested in. Professors can also refer to this mean when scaling their class averages.

Target Population, Parameters of Interest, and Sampling Details

We downloaded the 2021 winter session data set from this GitHub repo. We chose 2021W because (at the time of choosing) it was the most recent academic session available on GitHub. So, the target population was all courses in the 2021W session. Note that we filtered the dataset such that, for courses with multiple sections, we retained only the row labelled "overall" (i.e., the overall number of students, the overall average, etc). The parameters of interest were the mean course average and the population proportion of students with grades above 90%.

To determine the sample size required, we used a guess of the standard error of the sample. Since we didn't have access to the population standard deviation to guess the standard error of the sample, we decided to use our personal transcript standard deviations to estimate the sample standard error. We found our transcript standard deviations to be around 7 percent, so we decided

to guess the sample standard error to be 7 as well. This was not a perfect estimation (e.g., the grades from the dataset were from multiple students, and ours are from individuals), but it was adequate for this guess. Using this guess, we calculated a sample size for the mean and the proportion and chose the larger of the two ("n") as our actual sample size.

We obtained an SRS for the average by randomly choosing "n" courses from the dataset. To calculate the sample mean, we used the mean function in R; to calculate the sample proportion above 90%, we divided the number of course averages above 90% by n.

When choosing strata, we reasoned that there were far too many departments (probably more than 80) for us to stratify by department; so, we grouped departments into strata by field: arts, science, business, engineering, and other. Using a CSV file containing a table of departments and strata, we grouped courses (via their department) into strata (arts, science, business, engineering, other). We thought that the predicted variance would be the same for each stratum. We chose proportional allocation because proportional allocation is the optimal allocation when the predicted variance is the same for each stratum, and the cost of sampling from each stratum is the same.

Estimates, Standard Errors, and Confidence Intervals

From a simple random sample of 353 observations, we have estimated the mean of all course averages to be 81.90%, with a standard error of 0.37%. A 95% confidence interval for this estimate is [81.17%, 82.63%]. This means that we are 95% confident that the population mean of all course averages is between 81.17% and 82.63%. We assumed that the sample size (353) is large enough for the CLT to hold. The advantage of this method is that it is very simple to implement and very easy to intuitively understand the results. The disadvantage is that it has a relatively high standard error, and so is not very precise.

From the same simple random sample of 353 observations, we have estimated the proportion of course averages above 90% to be 0.156, with a standard error of 0.018. A 95% confidence interval for this estimate is [0.120, 0.192]. This means that we are 95% confident that the population proportion of course averages above 90% is between 0.120 and 0.192. We assumed that the sample size (353), the estimated proportion (0.156), and one minus the estimated proportion (0.844) are all large enough for the normal approximation of the binomial distribution to hold. The advantages and disadvantages of this method are the same as the SRS for the mean (above).

From a stratified sample of 353 observations, stratifying by field, and using proportional allocation, we have estimated the mean of all course averages to be 82.68%, with a standard error of 0.35%. A 95% confidence interval for this estimate is [81.99%, 83.36%]. This means

that we are 95% confident that the population mean of all course averages is between 81.99% and 83.36%. We assumed that the sample sizes for each stratum were large enough for the CLT to hold. The advantage of this method is that it has less standard error than an SRS (since variation from stratified sampling is only the within-strata variation). The disadvantage of this method is that it requires more effort to sample, as we need to collect data from each stratum.

From the same stratified sample of 353 observations, we have estimated the proportion of course averages above 90% to be 0.116, with a standard error of 0.016. A 95% confidence interval for this estimate is [0.084, 0.148]. This means that we are 95% confident that the population proportion of course averages above 90% is between 0.084 and 0.148. We assumed that the sample sizes, estimated proportions, and one minus the estimated proportions for each stratum are large enough for the normal approximation of the binomial distribution to hold. The advantages and disadvantages of this method are the same as the stratified sample for the mean (above).

Final Conclusions and Discussion

In 2021W, the estimated mean of the mean course average was (by either SRS or proportional stratified sampling) an A- grade. The estimated proportion of students above 90% was (by SRS) 0.156, with a standard error of 0.018, or (by stratified sampling) 0.116, with a standard error of 0.016.

A limitation of the conclusions is that it's based on random samples of the data, which is susceptible to the effects of randomness. Since we did not sample every observation in the population, the estimates and standard errors we obtained are subject to random variation. These conclusions cannot be reliably generalized to other populations, such as course averages at other institutions (since they may have different school-wide standards for grading), or even past or future years at UBC (since the standards for grading at UBC have always been evolving throughout the years).