2014-2015 PROGRAM EXIT ASSESSMENT

RE-REGISTRATION RATE (SAME COLLEGE, DIFFERENT PROGRAM) FOR COMPUTER SCIENCE STUDENTS

Matthew Schlachter
Computer Sciences
CÉGEP Heritage College
Gatineau (Hull), QC / J8Y 6T3
mschlachter@cegep-heritage.qc.ca

Key Words: Computer Science; Re-Registration Rate.

1. DATA CONTEXT

What the Re-registration rate data set is measuring is the percentage of computer science students re-registering to a different program in the same college by the cohort in their third semester. *Where* the data set was collected was at Heritage College and also the network (all public colleges in SRAM). *When* the data set was collected was from Fall 2004 through Fall 2013. *Why* the data was recorded was to keep historic data on students registering to the college, and was later aggregated to see trends in re-registration rates over time. *How* the data set was created was by aggregating data about individual students' registrations already stored in SRAM.

Population A consists of students who are new to collegial level studies, meaning students who are transfering directly from high school in Quebec or have otherwise never engaged in post-secondary education prior to entering computer science. Population B consists of students who have previous college studies in a different program and thus have transferred to Computer Science, who attended high school outside of Quebec, or who took a break from their studies before starting college. Within these two groups of students, they are split again in two ways: the year they began their studies in computer science and whether they went exclusively to Heritage or else went to a college in the network. Thus the *who* of the data is the individual students studying computer science at a college within the network, grouped by starting semester and whether or not they went to Heritage College.

2. DATA

These tables show the frequencies and totals for the different cohorts and populations, seperated between Heritage and the Network.

Table 1: Heritage Re-Registration Rate

Cohort	Population A	Population B	Total
F-2004	6	N/A	6
F-2005	6	3	9
F-2006	7	1	8
F-2007	4	0	4
F-2008	9	2	11
F-2009	8	3	11
F-2010	9	2	11
F-2011	10	4	14
F-2012	15	4	19
Total	74	19	93

Table 2: Network Re-Registration Rate

Cohort	Population A	Population B	Total
F-2004	571	334	905
F-2005	515	312	827
F-2006	524	284	808
F-2007	523	295	818
F-2008	670	380	1050
F-2009	712	353	1065
F-2010	616	384	1000
F-2011	670	420	1090
F-2012	686	427	1113
Total	5487	3189	8676

These barcharts show the relative re-registration rates per cohort, comparing Heritage to the Network and grouping by population.

Figure 1: Population A

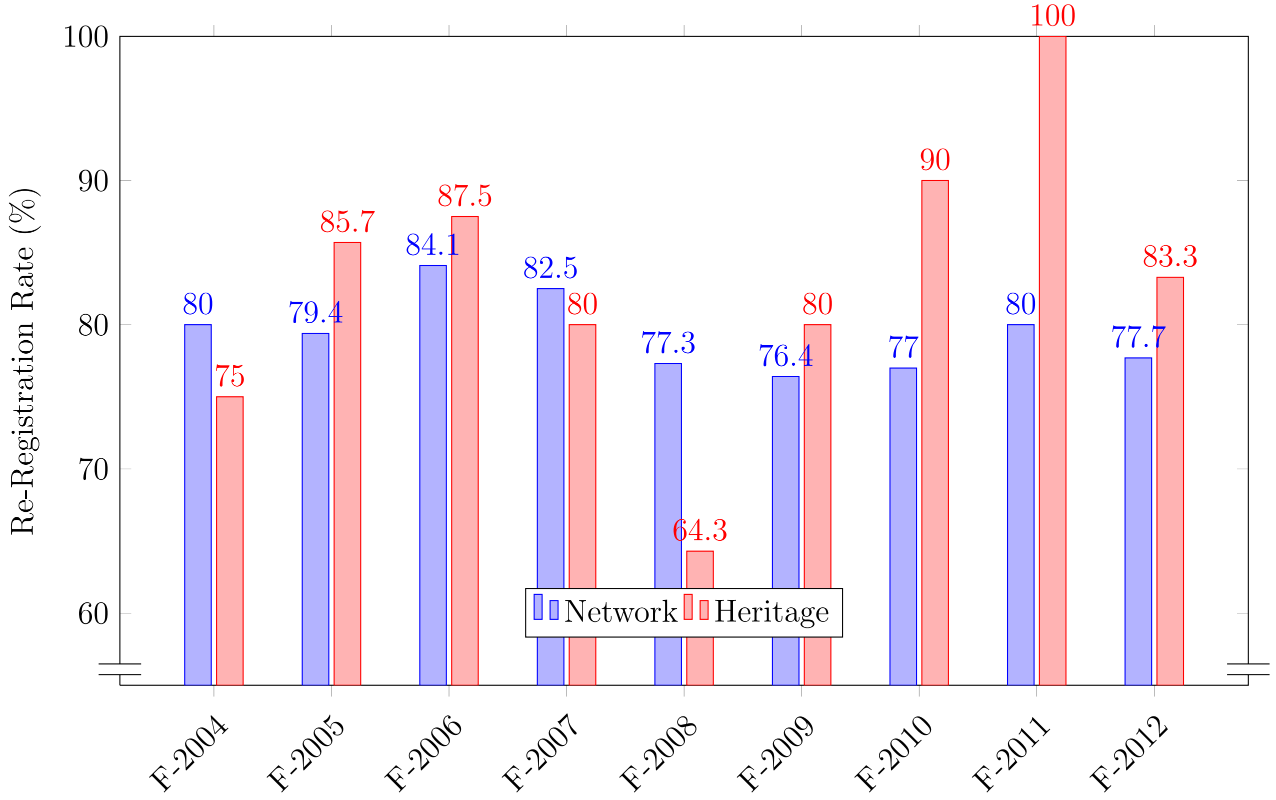
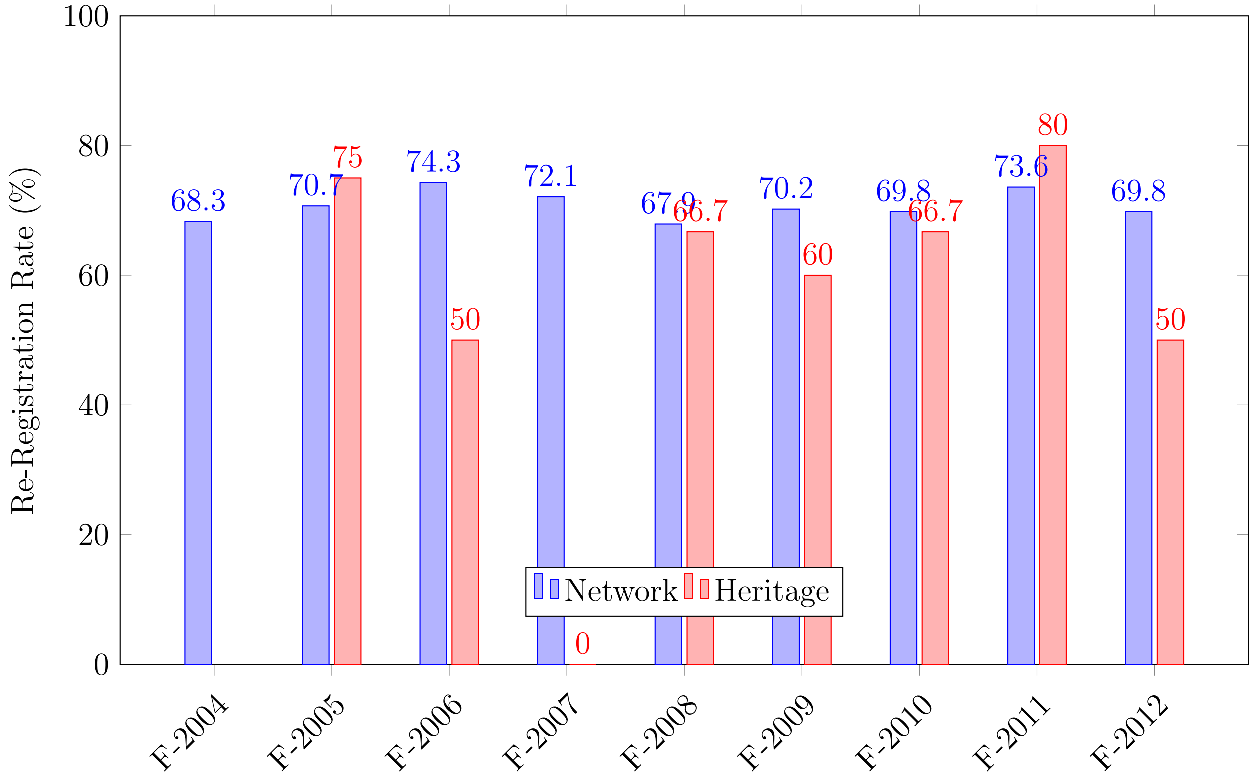


Figure 2: Population B



2. DATA ANALYSIS

2.1. 5-Number Summaries

Below are 5-number summaries for different populations based on their re-registration rates (in percentages), with each cohort being treated as a datapoint. These 5-number summeries include, in order: the minimum value, the first quartile value, the median value, the third quartile value, and the maximum value, which divide the data into quarters.

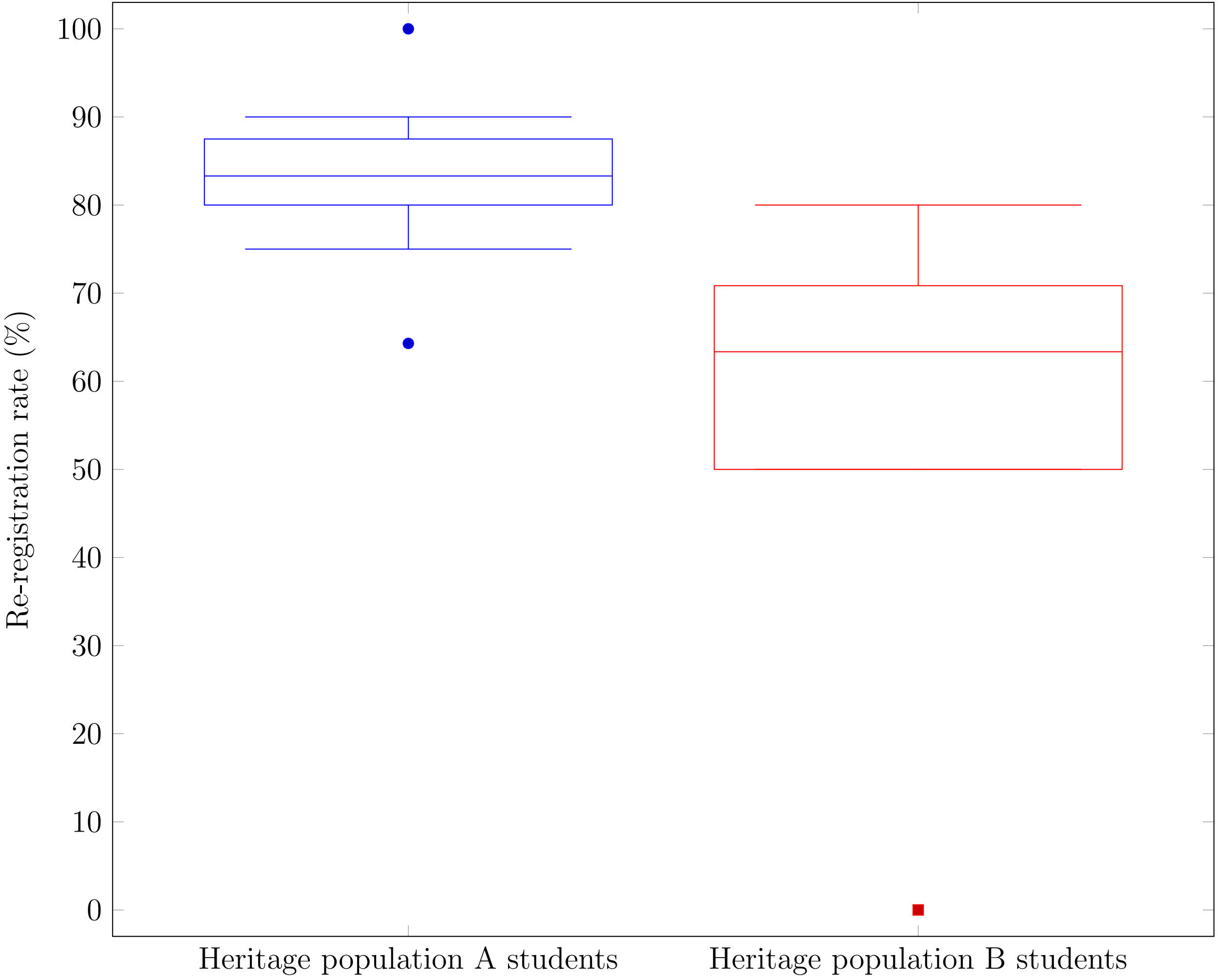
Table 3: 5-Number Summaries

Population	Min	Q1	Median	Q3	Max
Heritage population A	64.30	77.50	83.30	88.75	100.00
Heritage population B	0.00	50.00	63.35	70.85	80.00
Heritage	0.00	62.15	75.00	84.50	100.00
Network population A	76.40	77.15	79.40	81.25	84.10
Network population B	67.90	69.05	70.20	72.85	74.30
Network	67.90	70.20	75.35	79.40	84.10

2.2. Boxplots

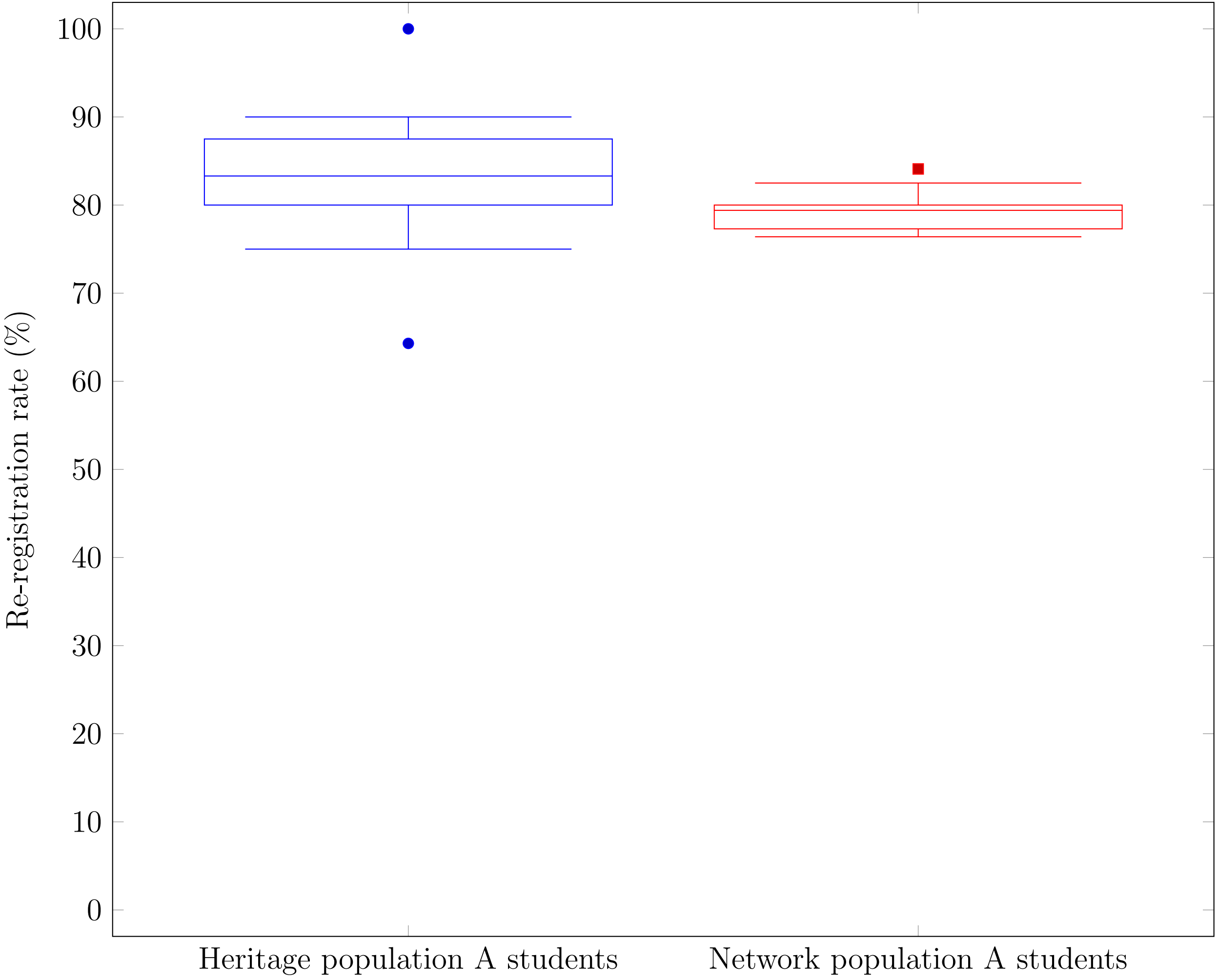
Below are pairs of boxplots displaying information about different populations. Each pair is followed by a comparison of the two boxplots.

Figure 3: Heritage Student Boxplots



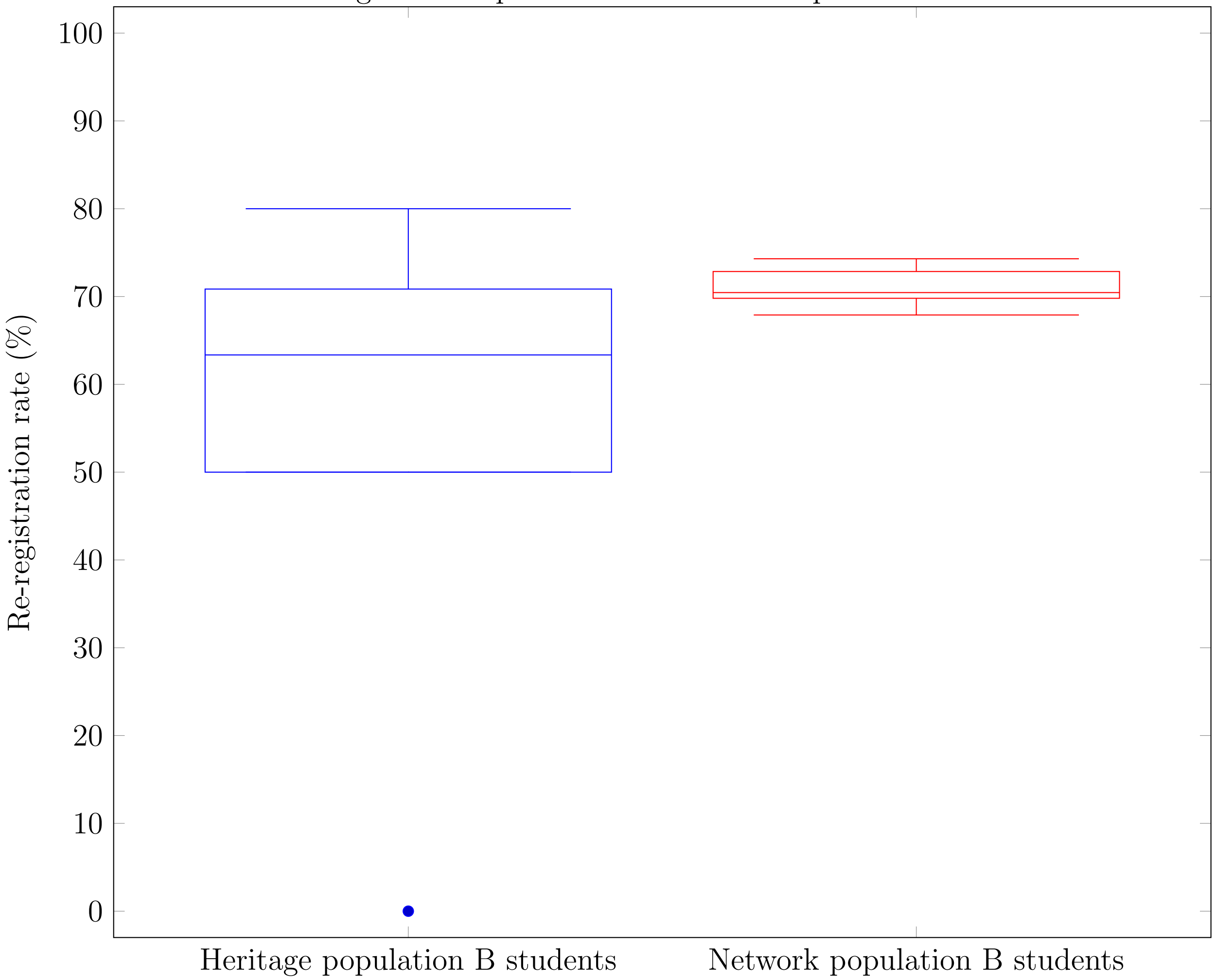
Population A is more symmetric than population B and has a much smaller range and inter-quartile range. Population A has twice as many outliers as population B, although the outliers for population A are much closer to the median than is the outlier of population B. About 75% of the cohorts in population A have a higher re-registration rate than any cohort in population B, since the first quartile of population A is near the upper whisker of population B. Population A had the highest re-registration rate for a cohort, while population B had the lowest re-registration rate for a cohort.

Figure 4: Population A Student Boxplots



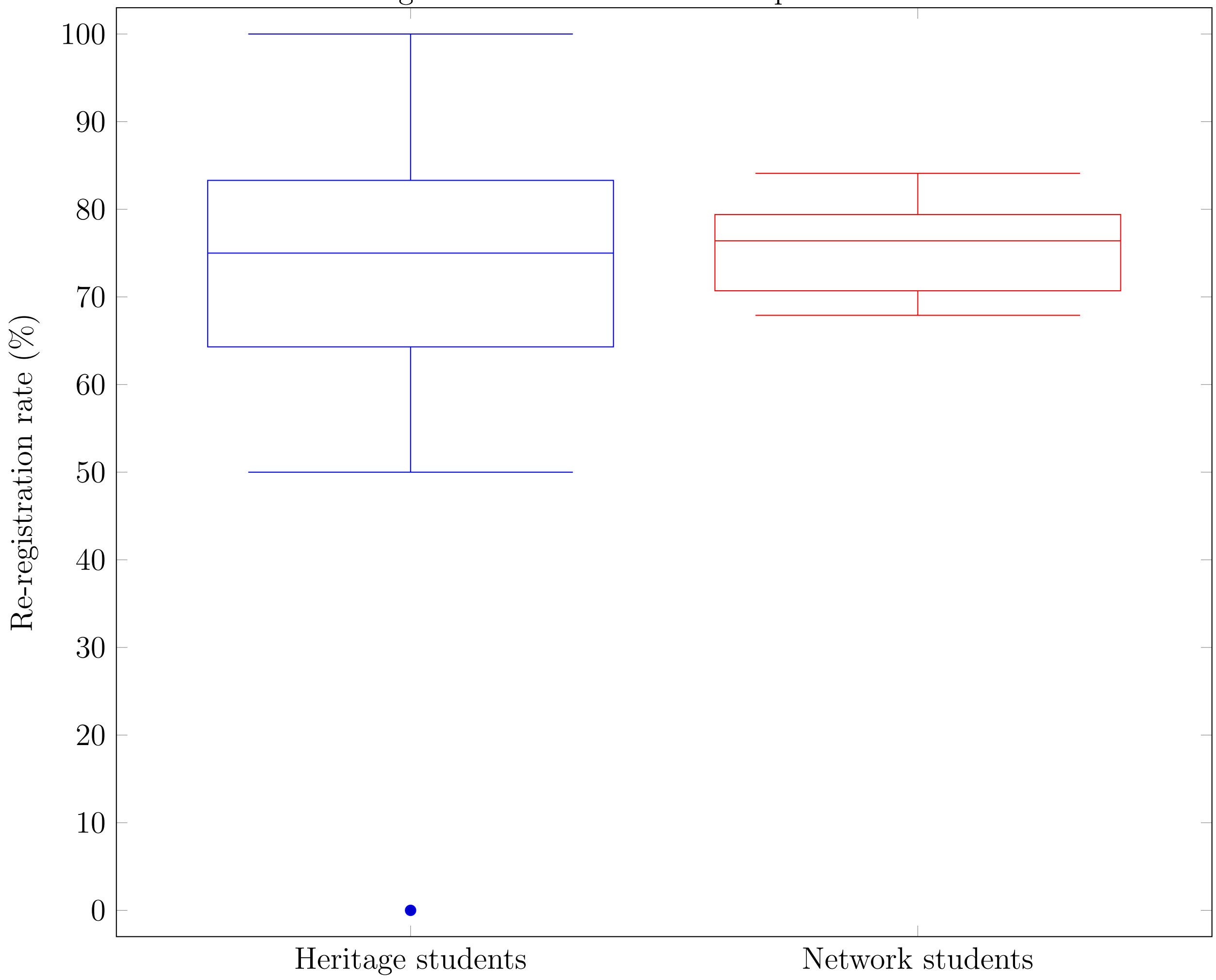
The Heritage population has both the highest range and the highest inter-quartile range. The Heritage population is more symmetric than the Network population, which seems to have a skew to the right. The Heritage has two outliers, which are essentially equidistant from its median, while the Network population only has one outlier, which is near the median of the Heritage population. About 75% of the Heritage cohorts have a higher re-registration rate than about 75% of the Network cohorts. The Heritage population had both the highest and the lowest re-registration rate for a cohort.

Figure 5: Population B Student Boxplots



The Heritage population has the highest range and the highest inter-quartile range. The Heritage population has the cohorts with both the highest and lowest re-registration rates. The Network population is more symmetric, although both are observably skewed; the Heritage population being skewed to the left (with the outlier included) and the Network population being slightly less skewed to the right. The re-registration rates for the Network population are somewhat tightly clustered around the third quartile of the Heritage population.

Figure 6: Overall Student Boxplots



The Heritage population has the highest range and the highest inter-quartile range. The Heritage population also has the cohorts with both the highest and lowest re-registration rates, and includes both the minimum and maximum possible values. The Heritage population and the Network population have about the same shape and symmetry (excluding the outlier). The Heritage population has one outlier while the Network population has none. The re-registration rates for the Network population nearly all fit into the middle 50% of the re-registration rates for the Heritage population.

3. CALCULATIONS

3.1 Sample Population Calculations

These calculations describe different populations, treating each as a single sample weighted by enrolment. Each set of calculations includes the sample mean (denoted $\overline{x}$), standard error (denoted $SE(\overline{x})$), and a 95% confidence interval (denoted CI). These calculations will use a total population mean of $p = 0.7573$ taken from the network population, and a z-score of 1.96 for the critical value (z^*) of the confidence interval. After each population, it will be discussed whether the neccessary conditions are met for a valid confidence interval.

$$\begin{aligned}\overline{x} &= \frac{x}{n} \qquad\qquad \text{(Calculations for Heritage College population A students)}\\ &= \frac{74}{90}\\ &= 0.8\overline{22}\end{aligned}$$

$$\begin{aligned}SE(\overline{x}) &= \sqrt{\frac{p(1-p)}{n}}\\ &= \sqrt{\frac{0.7573(0.2427)}{90}}\\ &\approx 0.0446\end{aligned}$$

$$\begin{aligned}CI &= \overline{x} \pm z^* * SE(\overline{x})\\ &= 0.8222 \pm 1.96 * 0.0446\\ &= 0.8222 \pm 0.0874\\ &= [0.7348, 0.9096]\end{aligned}$$

This sample satisfies the 10% condition since the sample size (90) is less than 10% of the total population (11,457). This sample satisfies the 10 success/10 failure condition since it is expected that there will be at least 10 successes and 10 failures ($np \geq 10$ and $n(1-p) \geq 10$). This sample is not random; however, we will assume it to be representative of the population. Thus, the confidence interval is valid.

$$\begin{aligned}\overline{x} &= \frac{x}{n} && \text{(Calculations for Heritage College population B students)} \\ &= \frac{19}{24} \\ &= 0.791\overline{66}\end{aligned}$$

$$\begin{aligned}SE(\overline{x}) &= \sqrt{\frac{p(1-p)}{n}} \\ &= \sqrt{\frac{0.7573(0.2427)}{24}} \\ &\approx 0.0864\end{aligned}$$

$$\begin{aligned}CI &= \overline{x} \pm z^* * SE(\overline{x}) \\ &= 0.7917 \pm 1.96 * 0.0864 \\ &= 0.7917 \pm 0.1693 \\ &= [0.6224, 0.9610]\end{aligned}$$

This sample satisfies the 10% condition since the sample size (24) is less than 10% of the total population (11,457). This sample does not satisfy the 10 success/10 failure condition since, although it is expected that there will be at least 10 successes, less than 10 failures are expected ($np \geq 10$ but $n(1-p) \leq 10$). This sample is not random; however, we will assume it to be representative of the population. Thus, the confidence interval is not valid.

$$\begin{aligned} \overline{x} &= \frac{x}{n} && \text{(Calculations for all Heritage College students)} \\ &= \frac{93}{114} \\ &\approx 0.8158 \end{aligned}$$

$$\begin{aligned} SE(\overline{x}) &= \sqrt{\frac{p(1-p)}{n}} \\ &= \sqrt{\frac{0.7573(0.2427)}{114}} \\ &\approx 0.0402 \end{aligned}$$

$$\begin{aligned} CI &= \overline{x} \pm z^* * SE(\overline{x}) \\ &= 0.8158 \pm 1.96 * 0.0402 \\ &= 0.8158 \pm 0.0788 \\ &= [0.7370, 0.8946] \end{aligned}$$

This sample satisfies the 10% condition since the sample size (114) is less than 10% of the total population (11,457). This sample satisfies the 10 success/10 failure condition since it is expected that there will be at least 10 successes and 10 failures ($np \geq 10$ and $n(1-p) \geq 10$). This sample is not random; however, we will assume it to be representative of the population. Thus, the confidence interval is valid.

3.2 Network Population Calculations and Hypotheses

$$\begin{aligned} p &= \frac{x}{n} && \text{(Calculations and hypotheses for population A Students)} \\ &= \frac{5487}{6940} \\ &\approx 0.7906 \end{aligned}$$

The null hypothesis for re-registration rate would be that the mean re-registration rate for Heritage students is the same as the mean for Network students such that:

$$H_0 : \overline{x} = 0.7906$$

An alternative hypotheses would be that the mean re-registration rate for Heritage students is greater than the mean for Network students such that:

$$H_A : \overline{x} > 0.7906$$

The independence condition is satisfied. The sample is not random so the randomization condition is not satisfied; however, we'll assume that the sample is representative of the population. The 10% condition is satisfied. The success/failure condition is satisfied.

$$\begin{aligned} z &= \frac{\overline{x} - p}{\sqrt{p(1-p)/n}} \\ z &= \frac{0.8222 - 0.7906}{\sqrt{0.7906(0.2094)/90}} \\ z &\approx 0.7368 \\ \therefore \textit{P-value} &\approx 0.22965 \end{aligned}$$

With a significance level of 10%, we fail to reject the null hypothesis since there is not evidence that Heritage students have a higher re-registration rate than Network students. The calculated P-value is the probability of seeing results as good as those of Heritage students or better, because of natural sampling variation. Since the calculated P-value is not less than the significance level of 10%, the null hypothesis cannot be rejected.

$$\begin{aligned} p &= \frac{x}{n} && \text{(Calculations and hypotheses for population B Students)} \\ &= \frac{3189}{4517} \\ &\approx 0.7060 \end{aligned}$$

The null hypothesis for re-registration rate would be that the mean re-registration rate for Heritage students is the same as the mean for Network students such that:

$$H_0 : \overline{x} = 0.7060$$

An alternative hypotheses would be that the mean re-registration rate for Heritage students is greater than the mean for Network students such that:

$$H_A : \overline{x} > 0.7060$$

The independence condition is satisfied. The sample is not random so the randomization condition is not satisfied; however, we'll assume that the sample is representative of the population. The 10% condition is satisfied. The success/failure condition is not satisfied (but we'll ignore that).

$$\begin{aligned} z &= \frac{\overline{x} - p}{\sqrt{p(1-p)/n}} \\ z &= \frac{0.7917 - 0.7060}{\sqrt{0.7060(0.2940)/24}} \\ z &\approx 0.9215 \\ \therefore \textit{P-value} &\approx 0.17879 \end{aligned}$$

With a significance level of 10%, we fail to reject the null hypothesis since there is not evidence that Heritage students have a higher re-registration rate than Network students. The calculated P-value is the probability of seeing results as good as those of Heritage students or better, because of natural sampling variation. Since the calculated P-value is not less than the significance level of 10%, the null hypothesis cannot be rejected.

$$\begin{aligned} p &= \frac{x}{n} && \text{(Calculations and hypotheses for all Students)} \\ &= \frac{8676}{11457} \\ &\approx 0.7573 \end{aligned}$$

The null hypothesis for re-registration rate would be that the mean re-registration rate for Heritage students is the same as the mean for Network students such that:

$$H_0 : \overline{x} = 0.7573$$

An alternative hypotheses would be that the mean re-registration rate for Heritage students is greater than the mean for Network students such that:

$$H_A : \overline{x} > 0.7573$$

The independence condition is satisfied. The sample is not random so the randomization condition is not satisfied; however, we'll assume that the sample is representative of the population. The 10% condition is satisfied. The success/failure condition is satisfied.

$$\begin{aligned} z &= \frac{\overline{x} - p}{\sqrt{p(1-p)/n}} \\ z &= \frac{0.8158 - 0.7573}{\sqrt{0.7573(0.2427)/114}} \\ z &\approx 1.4569 \\ \therefore \textit{P-value} &\approx 0.07215 \end{aligned}$$

With a significance level of 10%, we can reject the null hypothesis since there is evidence that Heritage students have a higher re-registration rate than Network students. The calculated P-value is the probability of seeing results as good as those of Heritage students or better, because of natural sampling variation. Since the calculated P-value is less than the significance level of 10%, the null hypothesis can be rejected in favour of the alternative hypothesis that Computer Science students at Heritage have a higher re-registration rate that that of students in the Network