Stat 257 Fall 2005
Sampling Rare and Sensitive Events

Example: A senior administrator at the University of Regina wishes to estimate the proportion of its students that have used cocaine, a sensitive subject. Students were classified into one of two strata–undergraduate and graduate–and were randomly sampled within the stratum. Since there was some concern that students might be unwilling to disclose their use of cocaine to a university official, the following *random response method* was used. The university official constructs a deck of 30 cards. On 26 of them are marked **N** for *never used cocaine* and 4 of them are marked **C** for *have used cocaine at least once*. Each sampled student was asked to draw a card from the deck and to respond *yes* if the letter agrees with the group that student belongs to. The results are as follows:

STRATA	total number of students	number sampled	number answering yes
undergraduate	8972	900	723
graduate	1548	150	117

(a) Construct a 95% confidence interval for the proportion of undergraduates who have used cocaine at least once.

(b) Construct a 95% confidence interval for the proportion of graduates who have used cocaine at least once.

(c) Based on your answers to **(a)** and **(b)**, is there a statistically significant difference in the use of cocaine between undergraduates and graduates? Is this surprising? Why or why not?

Solutions: **(a)** We find that for undergraduates, $n_1 = 723$, $n = 900$, $\theta = 4/30$. Thus, an estimator of p is given by

$$\hat{p} = \frac{n_1/n}{2\theta - 1} - \frac{1-\theta}{2\theta - 1} = \frac{723/900}{-22/30} - \frac{26/30}{-22/30} \approx 0.086.$$

The estimated variance is given by

$$s^2(\hat{p}) = \frac{1}{(2\theta-1)^2} \cdot \frac{1}{n} \cdot \frac{n_1}{n} \cdot \left(1 - \frac{n_1}{n}\right) = \frac{1}{(-22/30)^2} \cdot \frac{1}{900} \cdot \frac{723}{900} \cdot \left(1 - \frac{723}{900}\right) \approx 0.000326.$$

In other words, an approximate 95% confidence interval for p is given by $0.086 \pm 2(0.018)$.

(b) We find that for graduates, $n_1 = 117$, $n = 150$, $\theta = 4/30$. Thus, an estimator of p is given by

$$\hat{p} = \frac{n_1/n}{2\theta - 1} - \frac{1-\theta}{2\theta - 1} = \frac{117/150}{-22/30} - \frac{26/30}{-22/30} \approx 0.118.$$

The estimated variance is given by

$$s^2(\hat{p}) = \frac{1}{(2\theta-1)^2} \cdot \frac{1}{n} \cdot \frac{n_1}{n} \cdot \left(1 - \frac{n_1}{n}\right) = \frac{1}{(-22/30)^2} \cdot \frac{1}{150} \cdot \frac{117}{150} \cdot \left(1 - \frac{117}{150}\right) \approx 0.0021.$$

In other words, an approximate 95% confidence interval for p is given by $0.118 \pm 2(0.046)$.

(c) Since the confidence intervals computed from **(a)** and **(b)** overlap, there is no statistically significant difference in cocaine use between undergraduates and graduates. ETC.