

Twenty-six people took the exam. The average score was 61.0 and the standard deviation of the scores was 18.3. The range went from 32 to 98.

1. A department store is interested in customers' opinion about their new hours. From their list of customers who have store credit cards they randomly select a subset of customers to call on the phone. (They have phone numbers for their credit card holders.) Describe the target population, sampling frame and sampling unit for this survey. Are there any possible sources of selection bias?

**ans** The target population is all customers while the sampling frame is only those customers with credit cards. The sampling unit is a credit card customer with a phone. A possible source of selection bias is that customers with credit cards could be quite different than those without credit cards.

2. Suppose we wish to find a conservative 95% confidence interval for the proportion of recipes in a cook book that are vegetarian. We plan to take a srsWOR of the  $N = 2351$  recipes in the book. We want the length of our interval to be at most 0.06. Determine the minimum sample size needed to obtain such an interval.

**ans** From class

$$n = \frac{n_0}{1 + n_0/N} \quad \text{where} \quad n_0 = (1.96/(2 \times 0.03))^2 = 1067$$

So

$$n = \frac{1067}{1 + 1067/2351} = 733.8$$

or we take  $n = 734$ .

3. Suppose a county contains 10,000 single-resident houses where 2,000 have electric heat and 8,000 have nonelectric heat. A stratified sample of 100 houses was taken and the January electricity consumption (in kilowatt-hours) was recorded. The results are given below.

Type of heating	Number of houses	Average energy consumption	Sample variance
Electric	25	900	200,000
Nonelectric	75	400	90,000

a) Find a 95% confidence interval for the average number of kilowatt-hours used by houses in this county.

**ans**

$$\bar{y}_{str} = \frac{2000}{10000}900 + \frac{8000}{10000}400 = 500$$

$$\begin{aligned} \hat{V}(\bar{y}_{str}) &= \sum_h (1 - n_h/N_h) \left(\frac{n_h}{N_h}\right)^2 \frac{s_h^2}{n_h} \\ &= (1 - 25/2000)(1/5)^2(200000/25) + (1 - 75/8000)(4/5)^2(90000/75) \\ &= 316 + 760.8 = 1076.8 \end{aligned}$$

So the confidence interval is  $500 \pm 1.96\sqrt{1076.8}$  which is  $500 \pm 64.3$ .

b) Assuming the observed sample variances were the true population variances what would have been the optimal allocation of the 100 sampled units?

**ans** Recall optimal allocation happens when  $n_h \propto N_h \sigma_h$ . We find that

$$n_1 = \frac{2,000 \times \sqrt{200,000}}{2,000 \times \sqrt{200,000} + 8,000 \times \sqrt{90,000}} 100 = .27(100) = 27$$

and so we should have sample 27 Electric and 73 Nonelectric.

4. A fitness corporation with a total of 260,000 members in 200 different facilities around the country wishes to estimate the total number of its members that regularly play squash on its courts. It takes a random sample of 20 of its 200 facilities and for each selected facility it determines how many members used the squash courts at least 9 times in the past 6 months. For  $i = 1, 2, \dots, 20$  let  $y_i$  be the number of regular squash players and  $x_i$  be the number of members of the facility for the  $i$ th facility in the sample. Suppose

$$\bar{y}_{smp} = 100 \quad \bar{x}_{smp} = 1500 \quad \sum_{i=1}^{20} (y_i - x_i/15)^2 = 1,874,656$$

a) Find the approximate 95% confidence interval (based on the Ratio estimator) for the total number of squash players.

**ans** The point estimate is

$$N\hat{R}\bar{X} = 200 \times 1/15 \times 1300 = 17,333$$

The estimated variance of the estimator is

$$N^2 \frac{1-f}{n} \sum_{i=1}^{20} (y_i - x_i/15)^2 / 19 = 200^2 \frac{0.9}{20} \frac{1,874,656}{19}$$

So an approximate 95% confidence interval is the estimate plus and minus 1.96 times the square root of its estimated variance.

b) In the  $(x, y)$  plane make a rough sketch of what a plot of the sample values should look like to ensure that the estimate you calculated in part a) is better than the standard estimate of the total number of squash players which just uses the values of  $y_{smp}$ . Make sure to specify the scale of the  $x$  and  $y$  axes carefully.

**ans** In the  $(x, y)$  plane the points should lay along a line through the origin with increasing variability as  $x$  increases.

5. A population consists of  $N_m$  men and  $N_w$  women. Let  $N = N_m + N_w$ . To get a sample of  $n$  men we proceed as follows. Draw a person at random from the population. If they are a man he becomes the first member of the sample. If they are a woman they are set aside. In either case we next draw another person at random from the remaining  $N - 1$  people  $N - 1$ . We continue in this way until we have selected  $n$  men. Let Sam be a specified member of the population.

a) What is the probability that Sam is chosen on the first draw?

**ans**  $1/N$ .

b) Given that the first draw was a man what is the conditional probability that it was Sam?

**ans**

$$P(\text{sam}|\text{man}) = P(\text{sam and man})/P(\text{man}) = P(\text{sam})/p(\text{man}) = (1/N)/(N_m/N) = 1/N_m$$

c) Given that that the first man selected was on the second draw what is the conditional probability it was Sam?

**ans**

$$P(\text{woman then sam} | \text{woman then man}) = ((N_w/N)(1/(N-1))) / ((N_w/N)(N_m/(N-1))) = 1/N_m$$

d) What is the probability that Sam was the first man selected?

**ans** In the same way it is easy to show that given that the first man was selected on the  $k$ th draw then the probability it was Sam is  $1/N_m$ . So this must be the unconditional probability as well.