Stat 257.01 Fall 2004
Assignment #6 (Selected) Solutions

(2.7) (a) The **mean** is a good summary number for typical calories per serving for these drinks. As such, the **standard deviation** is a good summary number for the variation in the calories per serving.

(b) Since there are extreme values, the **median** is a good number summary for the typical cost per serving for these drinks. As such, the **inter-quartile range** (IQR=Q3−Q1) is a good summary number for the variation in the typical cost per serving.

(c) No, the total of the calories column does not provide useful information, for practical purposes. Similarly, the total of the cost column does not provide useful information, for practical purposes.

(d) By eliminating Hydra Fuel from the list, there is not much impact on the average calories per serving, the standard deviation is increased, the average cost per serving is slightly decreased, and the standard deviation of the cost per serving is decreased.

(e) In fact, there is no most influential drink on the average calories per serving. They all seem to contribute about equally.

(4.40) (a) An estimate of the true proportion of women preferring Brand A is given by

$$\hat{p} = \frac{1}{n}\sum y_i = \frac{65}{100} = 0.65.$$

A bound on the error of estimation (ignoring the fpc) is given by

$$B = 2\sqrt{\hat{V}(\hat{p})} = 2\sqrt{\frac{\hat{p}\hat{q}}{n-1}} = 2\sqrt{\frac{0.65 \cdot 0.35}{100-1}} \approx 0.096.$$

In other words, an approximate 95% confidence interval for p is given by 0.65 ± 0.096.

(b) The target population is the proportion of potential customers preferring Brand A.

(c) No, they certainly did not select a simple random sample.

(d) One should always be skeptical when analyzing the results of such a survey. A potential bias that immediately comes to mind is that people who stop by the booth may have a predisposition to Brand A, and are therefore more likely to stop. Another scenario to consider is that many people (who might otherwise participate in a survey) simply are not inclined to stop at booths in shopping malls.

(4.41) From the problem as stated, we find that $n = 64$, $\hat{\mu} = 18300$, and $s = 400$. Hence, an approximate bound on the error of estimation of $\hat{\mu}$ is given by

$$B = 2\sqrt{\hat{V}(\hat{\mu})} = 2\sqrt{\frac{s^2}{n}} = 2\sqrt{\frac{400^2}{64}} = 100.$$

Thus, an approximate 95% confidence interval for the true salary is given by $18300 \pm 100 = (18200, 18400)$. Since this interval does not contain 20100, there is statistical evidence to support the claim that these secretaries are being paid low wages. An important assumption is that the sample mean has a normal distribution. (Although, this is not always exactly reasonable, by taking enough sample data points, and appealing to the Central Limit Theorem, a normal distribution is often approximately correct.)

(5.17) We solve this problem by using the *cumulative frequency* scheme for optimally choosing strata as in section 5.9. We have

number of employees	frequency	$\sqrt{\text{frequency}}$	cumulative $\sqrt{\text{frequency}}$
0-10	2	1.41	1.41
11-20	4	2.00	3.41
21-30	6	2.45	5.86
31-40	6	2.45	8.31
41-50	5	2.24	10.55
51-60	8	2.83	13.38
61-70	10	3.16	15.64
71-80	14	3.74	20.28
81-90	19	4.36	24.64
91-100	13	3.61	28.25
101-110	3	1.73	29.98
111-120	7	2.65	32.62

Since we seek $L = 4$ strata, we find $32.62/4 \approx 8.155$, so that the stratum boundaries should be as close as possible to 8.155, 16.312, and 24.468. Thus, choose stratum boundaries at 8.32, 16.54, and 24.64. This puts 0-40 employees in stratum 1, 41-70 employees in stratum 2, 71-90 employees in stratum 3, and 91-120 employees in stratum 3.

(5.21) (a) The proportion of defectives in the lot is

$$\hat{p} = \frac{1}{n}\sum y_1 = \frac{6+10}{100} = 0.16$$

and a bound on the error of estimation (ignoring the fpc) is

$$B = 2\sqrt{\hat{V}(\hat{p})} = 2\sqrt{\frac{\hat{p}\hat{q}}{n-1}} = 2\sqrt{\frac{0.16 \cdot 0.84}{100-1}} \approx 0.074.$$

In other words, an approximate 95% confidence interval for p is given by 0.16 ± 0.074.

(b) In this case, we find that

$$\hat{p}_{\text{st}} = \frac{N_1}{N}\hat{p}_1 + \frac{N_2}{N}\hat{p}_2 = 0.6 \cdot \frac{6}{38} + 0.4 \cdot \frac{10}{62} \approx 0.159$$

and a bound on the error of estimation (ignoring the fpc) is

$$\begin{aligned} B = 2\sqrt{\hat{V}(\hat{p}_{\text{st}})} &= 2\sqrt{\frac{N_1^2}{N^2} \cdot \frac{\hat{p}_1\hat{q}_1}{n_1 - 1} + \frac{N_2^2}{N^2} \cdot \frac{\hat{p}_2\hat{q}_2}{n_2 - 1}} \\ &= 2\sqrt{0.6^2 \cdot \frac{6/38 \cdot 32/38}{38-1} + 0.4^2 \cdot \frac{10/62 \cdot 52/62}{62-1}} \\ &\approx 0.081. \end{aligned}$$

In other words, an approximate 95% confidence interval for p_{st} is given by 0.159 ± 0.081.

Even though both **(a)** and **(b)** yield roughly the same "answer," I think that (b) is more acceptable because the quality inspector has *a priori* knowledge of the natural strata including the percentage of chips made by each assembly operation.