Problem Set #2

1a. The regular population variance can be given as

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Plugging in our new mean:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{\#1}{N}\right)^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i^2 - 2x_i\left(\frac{\#1}{N}\right) + \left(\frac{\#1}{N}\right)^2\right)$$

But $\sum x_i^2 = \sum x_i$ since each value of x will either be one or zero, which is the same squared.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}x_i^2 - 2\left(\frac{1}{N}\sum_{i=1}^{N}x_i * \sum_{i=1}^{N}\mu\right) + \frac{1}{N}\sum_{i=1}^{N}\mu^2 = \mu - \frac{2\mu}{N} * N\mu + \frac{1}{N} * N\mu^2$$

$$\sigma^2 = \mu - \frac{2\mu^2 N}{N} + \frac{N\mu^2}{N} = \mu - 2\mu^2 + \mu^2 = \mu - \mu^2$$

$$\sigma^2 = \mu(1 - \mu)$$

1b. We have the unbiased estimate for population variance is

$$\tilde{s}^2 = \left(1 - \frac{1}{N}\right) * \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$$

$$= \frac{N-1}{Nn-N}\left(\sum_{i=1}^{n}X_i^2 - 2\overline{X_n}\sum_{i=1}^{n}X_i + \sum_{i=1}^{n}\overline{X_n}^2\right)$$

Multiply by 1/n and n:

$$= \frac{Nn-n}{Nn-N}\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2 - \frac{2\overline{X_n}1}{n}\sum_{i=1}^{n}X_i + \frac{1}{n}\sum_{i=1}^{n}\overline{X_n}^2\right)$$

$$= \frac{Nn-n}{Nn-N}\left(\overline{X_n} - 2\overline{X_n}^2 + \overline{X_n}^2\right)$$

$$= \left(\overline{X_n} - \overline{X_n}^2\right)\frac{Nn-n}{Nn-N} = \overline{X_n}(1 - \overline{X_n})\frac{Nn-n}{Nn-N}$$

1c. We have N = 300, n = 90, and α = 0.05. So, $\overline{X_n}$ = 70/90:

$$I = \overline{X_n} \pm z_{1-\frac{\alpha}{2}} se[\overline{X_n}] = \frac{7}{9} \pm z_{0.975} se[\overline{X_n}]$$

$z_{0.975} = 1.96$ and:

$$se[\overline{X_n}] = \frac{s}{\sqrt{n}}\sqrt{1-\frac{n-1}{N-1}} = \frac{\sqrt{\overline{X_n}(1-\overline{X_n}) * \frac{Nn-n}{Nn-N}}}{\sqrt{n}}\sqrt{1-\frac{n-1}{N-1}}$$

$$= \frac{\sqrt{\frac{7}{9}\left(1-\frac{7}{9}\right) * \frac{300*90-90}{300*90-300}}}{\sqrt{90}}\sqrt{1-\frac{90-1}{300-1}} = 0.037$$

Our interval is:

$$I = \frac{7}{9} \pm 1.96 * 0.037 = (0.706, 0.850)$$

2. Since this sampling takes the same form as Question 1 (0 = No, 1 = Yes), we can use the population variance that we calculated:

$$\sigma^2 = \mu(1-\mu)$$

Since this is the worst case, we will ignore the statistic and focus on the margin of error. This will make sure that the interval is at most four percentage points wide.

$$I = \overline{X_n} \pm z_{1-\frac{\alpha}{2}} se[\overline{X_n}]$$

$$z_{1-\frac{\alpha}{2}} se[\overline{X_n}] \leq 0.02$$

$$1.96 * \frac{\sqrt{\frac{n}{N}\left(1-\frac{n}{N}\right) * \frac{(N-1)n}{N(n-1)}}}{\sqrt{n}}\sqrt{1-\frac{n-1}{N-1}} \leq 0.02$$

Then, we ignore the finite population correction and assume worst-case scenario for variance. Since the population and sample sizes are large, $s^2 \approx \sigma^2$:

$$1.96 * \frac{\sqrt{\overline{X_n}(1-\overline{X_n})}}{\sqrt{n}} \leq 0.02$$

$$1.96 * \frac{\sqrt{0.5(1-0.5)}}{\sqrt{n}} \leq 0.02$$

$$n \geq 2401$$

The minimum sample size is 2401 students.

Problem 3 is attached as code.

4a. We know that

$$bias\left[\hat{\theta}_n\right] = E\left[\hat{\theta}_n\right] - \theta$$

With $E[\bar{X}_n] = \frac{\theta}{2}$:

$$E\left[\hat{\theta}_n\right] = E[2\bar{X}_n] = 2E[\bar{X}_n] = 2\left(\frac{\theta}{2}\right) = \theta$$

$$bias\left[\hat{\theta}_n\right] = \theta - \theta = 0$$

This is an unbiased estimator. Finding se:

$$se\left[\hat{\theta}_n\right] = \sqrt{V\left[\hat{\theta}_n\right]} = \sqrt{V[2\bar{X}_n]} = \sqrt{4V[\,\bar{X}_n]} = \sqrt{\frac{4\sigma^2}{n}}$$

We can use this formula since the population size of the uniform distribution is infinite.

$$se\left[\hat{\theta}_n\right] = \sqrt{\frac{4\theta^2}{12n}} = \frac{\theta}{\sqrt{3n}}$$

Then, MSE:

$$MSE\left[\hat{\theta}_n\right] = bias\left[\hat{\theta}_n\right]^2 + se\left[\hat{\theta}_n\right]^2 = 0 + \frac{\theta^2}{3n} = \frac{\theta^2}{3n}$$

4b. Again,

$$bias\left[\hat{\theta}_n\right] = E\left[\hat{\theta}_n\right] - \theta$$

We have

$$bias\left[\hat{\theta}_n\right] = \mathrm{E}[\max\{X_1, \dots, X_n\}] - \theta$$

To find the expected value, we can create a PDF and integrate:

$$CDF = P(\max\{x_1, \dots, x_n\} \leq x) = P(x_1 \leq x, x_2 \leq x, \dots, x_n \leq x)$$

$$\prod_{i=1}^{n} P(x_i \leq x)$$

Which is any value between zero and x:

$$\prod_{i=1}^{n} \frac{x-0}{\theta} = \left(\frac{x}{\theta}\right)^n$$

We differentiate to get the PDF:

$$PDF = \frac{nx^{n-1}}{\theta^n}$$

Therefore, our expected value is

$$E\left[\hat{\theta}_n\right] = \int_0^{\theta} x * \left(\frac{nx^{n-1}}{\theta^n}\right) dx$$

$$\left[\frac{nx^{n+1}}{\theta^n(n+1)}\right]_0^{\theta} = \frac{n\theta^{n+1}}{\theta^n(n+1)} = \frac{n\theta}{n+1}$$

$$Bias\left[\hat{\theta}_n\right] = \frac{n\theta}{n+1} - \theta = -\frac{\theta}{n+1}$$

For the se, we need the variance:

$$V\left[\hat{\theta}_n\right] = E\left[\hat{\theta}_n^2\right] - E\left[\hat{\theta}_n\right]^2$$

Using the same logic as the bias:

$$E\left[\hat{\theta}_n^2\right] = \int_0^{\theta} x^2 * \left(\frac{nx^{n-1}}{\theta^n}\right) dx = \int_0^{\theta} \left(\frac{nx^{n+1}}{\theta^n}\right) dx$$

$$\left[\frac{nx^{n+2}}{\theta^n(n+2)}\right]_0^{\theta} = \frac{n\theta^2}{n+2}$$

$$se\left[\hat{\theta}_n\right] = \sqrt{\frac{n\theta^2}{n+2} - \left(\frac{n\theta}{n+1}\right)^2} = \sqrt{\frac{n\theta^2}{(n+2)(n+1)^2}}$$

MSE:

$$MSE\left[\hat{\theta}_n\right] = bias\left[\hat{\theta}_n\right]^2 + se\left[\hat{\theta}_n\right]^2 = \frac{\theta^2}{(n+1)^2} + \frac{n\theta^2}{(n+2)(n+1)^2}$$

$$= \frac{2\theta^2}{(n+2)(n+1)}$$

c. We see which estimate has a smaller MSE:

$$\frac{\theta^2}{3n} ? = \frac{\theta^2}{\frac{1}{2}(n+2)(n+1)}$$

The numerator is the same, so we have to see which denominator is bigger:

$$\frac{1}{2}(n+2)(n+1) - 3n > 0$$

Which is the same for n = 1 and n = 2, but $\hat{\theta}_n = \max\{X_1, \dots, X_n\}$ has the smaller MSE for n > 2. Therefore, $X_{(n)}$ is more efficient.