1 Estimating Proportions

In the context of this problem, we are given that the population mean μ is simply the proportion of 1s such that $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{\#1}{N}$.

1.1 Part a

Let us show that the population variance in this case is $\sigma^2 = \mu(1-\mu)$:

We know that population variance is generally defined as $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$. With this definition, we can being to arrive at our solution:

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

$$\implies \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \frac{1}{N}\sum_{i=1}^{N}2\cdot x_i\cdot\mu + \frac{1}{N}\sum_{i=1}^{N}\mu^2$$

$$\implies \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \frac{2}{N}\cdot\sum_{i=1}^{N}x_i\cdot\sum_{i=1}^{N}\mu + \frac{1}{N}\sum_{i=1}^{N}\mu^2$$

Due to the provided definition of x_i, we can clearly see that $\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} x_i^2$. This is because x_i is defined to be either 1 or 0, thus resulting in the same value if it is squared. With this, we can continue to arrive at our solution knowing that $\frac{1}{N}\sum_{i=1}^{N} x_i^2 = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{\#1}{n} = \mu$:

$$\implies \sigma^2 = \mu - \frac{2}{N}\cdot\mu\cdot N\mu + \frac{1}{N}\cdot N\cdot\mu^2$$

$$\implies \sigma^2 = \mu - 2\mu^2 + \mu^2 = u - u^2 = \mu(1-\mu)\ \checkmark$$

1.2 Part b

Let us use the result in (a) to derive an unbiased estimate $\tilde{s}^2$ of σ^2 that depends only on the sample mean $\bar{X}_n$, the sample size n, and the population size N:

From Survey Sampling II Corollary 1, we know that an unbiased estimate for the population variance σ^2 is

$$s^2 = (1-\frac{1}{N})\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

Let us use this definition to derive $\tilde{s}^2$ under the stated conditions:

$$\implies \tilde{s}^2 = (1 - \frac{1}{N})\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{N-1}{N} \cdot \frac{1}{n-1} \cdot \sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X}_n + \bar{X}_n^2)$$

$$\implies \tilde{s}^2 = \frac{N-1}{N(n-1)} \cdot (\sum_{i=1}^{n} X_i^2 - 2\bar{X}_n \cdot \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} \bar{X}_n^2)$$

Following the same logic as (a), we know $\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n$, so let us multiply by $\frac{n}{n}$ in order to properly utilize these implications:

$$\implies \tilde{s}^2 = \frac{n}{n} \cdot \frac{N-1}{N(n-1)} \cdot (\sum_{i=1}^{n} X_i^2 - 2\bar{X}_n \cdot \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} \bar{X}_n^2)$$

$$\implies \tilde{s}^2 = \frac{n(N-1)}{N(n-1)} \cdot \frac{1}{n}(\sum_{i=1}^{n} X_i^2 - 2\bar{X}_n \cdot \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} \bar{X}_n^2)$$

$$\implies \tilde{s}^2 = \frac{n(N-1)}{N(n-1)} \cdot (\frac{1}{n}\sum_{i=1}^{n} X_i^2 - 2\bar{X}_n \cdot \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n}\sum_{i=1}^{n} \bar{X}_n^2)$$

$$\implies \tilde{s}^2 = \frac{n(N-1)}{N(n-1)} \cdot (\bar{X}_n - 2\bar{X}_n \cdot \bar{X}_n + \bar{X}_n^2)$$

$$\implies \tilde{s}^2 = \frac{n(N-1)}{N(n-1)} \cdot (\bar{X}_n - \bar{X}_n^2)$$

1.3 Part c

Using the results from (b), let us construct a 95% confidence interval for the proportion of students in the class who like statistics:

From Survey Sampling III, we know that a $100(1-\alpha)\%$ confidence interval I is defined as

$$I = \bar{X}_n \pm z_{1-\frac{\alpha}{2}} \cdot se[\bar{X}_n]$$

We have $N = 300, n = 90, \alpha = 0.05, \bar{X}_n = \frac{70}{90}, z_{1-\frac{\alpha}{2}} = z_{0.975} = 1.96$. Also, we know $se[\bar{X}_n] = \sqrt{\mathbb{V}[\bar{X}_n]} = \sqrt{\frac{\sigma^2}{n} \cdot (1 - \frac{n-1}{N-1})}$ from Definition 1 and Theorem 2 of Survey Sampling I. Thus, we can solve as follows:

$$\implies I = \frac{70}{90} \pm 1.96 \cdot \sqrt{\frac{\tilde{s}^2}{n} \cdot (1 - \frac{n-1}{N-1})}$$

$$\implies I = \frac{70}{90} \pm 1.96 \cdot \sqrt{\frac{\frac{n(N-1)}{N(n-1)} \cdot (\bar{X}_n - \bar{X}_n^2)}{n} \cdot (1 - \frac{n-1}{N-1})}$$

$$\implies I = \frac{70}{90} \pm 1.96 \cdot \sqrt{\frac{\frac{90(300-1)}{300(90-1)} \cdot (\frac{70}{90} - (\frac{70}{90})^2)}{90} \cdot (1 - \frac{90-1}{300-1})}$$

$$\implies I = \frac{70}{90} \pm 1.96 \cdot 0.0369 = (0.706, 0.850)$$

2 Finding the Sample Size

Let us find the minimum sample size required to make a 95% confidence interval for the population percentage that is at most 4 percentage points wide (length of interval is at most 0.04). We will assume the worst-case scenario for variance and ignore the finite population correlation.

This survey has the same structure as the surveys explored in Problem 1, so let us utilize the population variance defined as $\sigma^2 = \mu(1-\mu)$. Again, from Survey Sampling III, we can define a $100(1-\alpha)\%$ confidence interval I is defined as $I = \bar{X}_n \pm z_{1-\frac{\alpha}{2}} \cdot se[\bar{X}_n]$. Knowing that the length of the interval is at most 0.04, we can arrive at the following implications:

$$I = \bar{X}_n \pm z_{1-\frac{\alpha}{2}} \cdot se[\bar{X}_n] \implies z_{1-\frac{\alpha}{2}} \cdot se[\bar{X}_n] \leq 0.02$$

Let us essentially solve for n with a 95% confidence interval.

$$\implies z_{0.975} \cdot se[\bar{X}_n] \leq 0.02$$

$$\implies 1.96 \cdot \sqrt{\frac{s^2}{n}} \leq 0.02 \quad \text{*ignoring finite population correction}$$

$$\implies 1.96 \cdot \sqrt{\frac{\sigma^2}{n}} \leq 0.02 \quad (s^2 \approx \sigma^2)$$

$$\implies 1.96 \cdot \sqrt{\frac{0.5(1-0.5)}{n}} \leq 0.02 \quad \text{*worst-case variance } \mu = 0.5$$

$$\implies \frac{\sqrt{0.5(1-0.5)}}{\sqrt{n}} \leq \frac{0.02}{1.96} \implies \sqrt{n} \geq \sqrt{0.5(1-0.5)} \cdot \frac{1.96}{0.02}$$

$$\implies n \geq 2401$$

Thus, the minimum sample size required is 2401 students.

3 The Bootstrap Method

See following Matlab file:

PS2 - Q3

Ben Juarez

Part a

```
birth = importdata("birth.txt");
bwt = birth(:,1);
bwt = bwt * 0.0283495;
bwt = bwt(bwt<999);
n = 100;
N = length(bwt);

pop_mean = mean(bwt);
disp("Population mean weight:");disp(pop_mean);
```

```
Population mean weight:
    3.3899
```

```
X = datasample(bwt, n);
sample_mean = mean(X);
disp("Sample mean weight:");disp(sample_mean);
```

```
Sample mean weight:
    3.4300
```

```
se = sqrt(((std(bwt)^2)/n) * (1 - ((n - 1)/(N - 1))));
disp("Standard error of sample mean:");disp(se);
```

```
Standard error of sample mean:
    0.0496
```

Part b

```
B = 10^3;
P = repmat(X, round(N/n), 1);
sample_means = zeros();
for i = 1:B
    sample_means(i) = mean(datasample(P,n));
end
sum = 0;
avg_sample_means = mean(sample_means);
for i = 1:B
    sum = sum + (sample_means(i) - avg_sample_means)^2;
end
se_hat = sqrt(1/B * sum);
disp("Bootstrap estimate of standard error:");disp(se_hat);
```

```
Bootstrap estimate of standard error:
    0.0519
```

Part c

```
k = floor(N/n);
```

```
r = rem(N, n);
p = (1 - (r/n))*(1-(r/(N-1)));
P1 = repmat(X, k, 1);
P2 = repmat(X, k+1, 1);
means = zeros();
for i = 1:B
    if rand() < p
        means(i) = mean(datasample(P1, n));
    else
        means(i) = mean(datasample(P2, n));
    end
end
sum = 0;
avg_sample_means = mean(means);
for i = 1:B
    sum = sum + (means(i) - avg_sample_means)^2;
end
se_hat = sqrt(1/B * sum);
disp("Bootstrap estimate of standard error:");disp(se_hat);
```

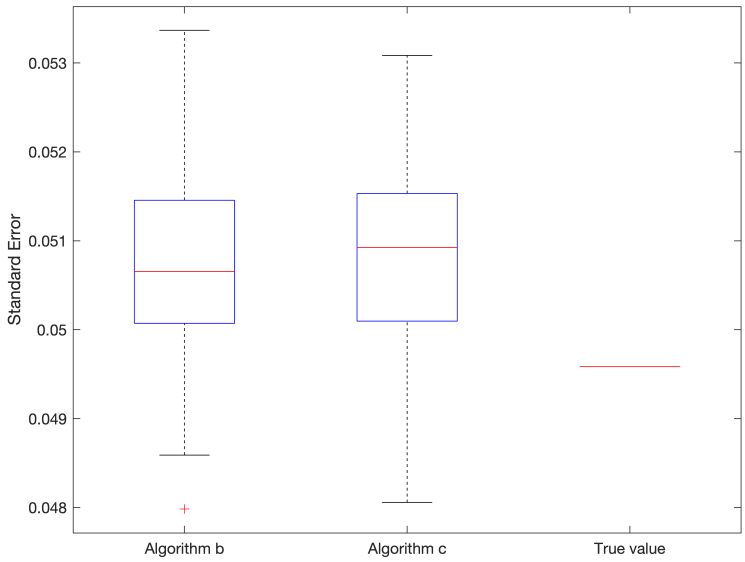
```
Bootstrap estimate of standard error:
    0.0507
```

Part d

```
part_b = zeros();
for z = 1:100
    P = repmat(X, round(N/n), 1);
    sample_means = zeros();
    for i = 1:B
        sample_means(i) = mean(datasample(P,n));
    end
    sum = 0;
    avg_sample_means = mean(sample_means);
    for i = 1:B
        sum = sum + (sample_means(i) - avg_sample_means)^2;
    end
    part_b(z) = sqrt(1/B * sum);
end
part_c = zeros();
for z = 1:100
    k = floor(N/n);
    r = rem(N, n);
    p = (1 - r/n)*(1-r/(N-1));
    P1 = repmat(X, k, 1);
    P2 = repmat(X, k+1, 1);
    means = zeros();
    for i = 1:B
        if rand() < p
            means(i) = mean(datasample(P1, n));
        else
            means(i) = mean(datasample(P2, n));
        end
    end
    sum = 0;
```

```
    avg_sample_means = mean(means);
    for i = 1:B
        sum = sum + (means(i) - avg_sample_means)^2;
    end
    part_c(z) = sqrt(1/B * sum);
end

figure;
boxplot([transpose(part_b), transpose(part_c), repmat(se, 100, 1)], ...
    'Labels', {'Algorithm b', 'Algorithm c', 'True value'});
ylabel("Standard Error");
```



Examining the box plot, we can see that both algorithms have fairly similar distributions. Additionally, the true value rarely falls within the IQR of either distribution. For most iterations, the true value falls within the largest/smallest observations (excluding outliers).

4 Bias, SE, and MSE

Suppose that $X_1, ..., X_n$ can be modeled as sample from uniform distribution on $[0, \theta]$ (where θ is unknown.

4.1 Part a

Let us find the bias, se, and MSE of this estimate with $\hat{\theta} = 2\bar{X}_n$:

For bias:

$$bias[\hat{\theta}_n] = \mathbb{E}[\hat{\theta}_n] - \theta \quad \text{*Modeling and Inference Lecture}$$

$$\implies \mathbb{E}[X_i] = \frac{\theta}{2}$$

$$\implies \mathbb{E}[\bar{X}_n] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\frac{\theta}{2} = \frac{1}{n}\cdot\frac{n\theta}{2} = \frac{\theta}{2}$$

$$\implies \mathbb{E}[\hat{\theta}] = \mathbb{E}[2\bar{X}_n] = 2\cdot\mathbb{E}[\bar{X}_n] = 2\cdot\frac{\theta}{2} = \theta \implies bias[2\bar{X}_n] = 0$$

For se:

$$se[\hat{\theta}_n] = \sqrt{\mathbb{V}[\hat{\theta}_n]} \quad \text{*unbiased estimator, ignore finite population correction}$$

$$\implies \mathbb{V}[X_i] = \frac{\theta^2}{12} \quad \text{*uniform distribution}$$

$$\implies \mathbb{V}[\bar{X}_n] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}[X_i] = \frac{1}{n^2}\sum_{i=1}^{n}\frac{\theta^2}{12} = \frac{1}{n^2}\cdot\frac{n\theta^2}{12} = \frac{\theta^2}{12n}$$

$$\mathbb{V}[\hat{\theta}] = \mathbb{V}[2\bar{X}_n] = 4\mathbb{V}[\bar{X}_n] = 4\cdot\frac{\theta^2}{12n} = \frac{\theta^2}{3n}$$

$$\implies se[2\bar{X}_n] = \sqrt{\mathbb{V}[2\bar{X}_n]} = \frac{\theta}{\sqrt{3n}}$$

For MSE:

$$MSE[\hat{\theta}_n] = bias[\hat{\theta}_n]^2 + se[\hat{\theta}_n]^2 \quad \text{*Modeling and Inference lecture}$$

$$\implies MSE[\hat{\theta}] = MSE[2\bar{X}_n] = (0)^2 + (\frac{\theta}{\sqrt{3n}})^2 = \frac{\theta^2}{3n}$$

4.2 Part b

Let us find the bias, se, and MSE of this estimate with $\hat{\theta} = X_{(n)} = max\{X_1, ..., X_n\}$:

For bias:

$$bias[\hat{\theta}_n] = \mathbb{E}[\hat{\theta}_n] - \theta \quad \text{*Modeling and Inference lecture}$$

$$\implies bias[\hat{\theta}] = bias[X_{(n)}] = \mathbb{E}[X_{(n)}] - \theta = \mathbb{E}[max\{X_1, ..., X_n\}] - \theta$$

In order to solve for $\mathbb{E}[X_{(n)}]$, we must find the PDF $f(x)$ (derived from the CDF $F(x)$) and then solve for

$$\mathbb{E}[X_{(n)}] = \int_0^\theta x \cdot f(x)dx$$

So, let us define the CDF $F(x)$ as follows:

$$\implies F(x) = \mathbb{P}(max\{X_1, ..., X_n\} \leq x) = \mathbb{P}(x_1 \leq x, x_2 \leq x, x_3 \leq x, ..., x_n \leq x)$$

$$\implies F(x) = \prod_{i=1}^{n} \mathbb{P}(x_i \leq x) = \prod_{i=1}^{n} \frac{x - 0}{\theta} = \frac{x^n}{\theta^n}$$

$$\implies f(x) = \frac{dF(x)}{dx} = \frac{nx^{n-1}}{\theta^n}$$

$$\implies \mathbb{E}[X_{(n)}] = \int_0^\theta x \cdot \frac{nx^{n-1}}{\theta^n}dx = \int_0^\theta = \frac{nx^n}{\theta^n}dx = \frac{nx^{n+1}}{(n+1)\theta^n}|_0^\theta = \frac{n\theta^{n+1}}{(n+1)\theta^n} = \frac{n\theta}{n+1}$$

$$\implies bias[X_{(n)}] = \frac{n\theta}{n+1} - \theta = \frac{n\theta}{n+1} - \frac{\theta(n+1)}{n+1} = -\frac{\theta}{n+1}$$

For se:

$$se[\hat{\theta}_n] = \sqrt{\mathbb{V}[\hat{\theta}_n]} \quad \text{*unbiased estimator, ignore finite population correction}$$

We know we can define the variance as follows:

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

So, we need to solve for $\mathbb{E}[\hat{\theta}^2] = \mathbb{E}[X_{(n)}^2]$. We arrive at the following:

$$\mathbb{E}[X_{(n)}^2] = \int_0^\theta x^2 \cdot \frac{nx^{n-1}}{\theta^n}dx \int_0^\theta = \frac{nx^{n+1}}{\theta^n}dx = \frac{nx^{n+2}}{(n+2)\theta^n}|_0^\theta = \frac{n\theta^{n+2}}{(n+2)\theta^n} = \frac{n\theta^2}{n+2}$$

$$\implies se[\hat{\theta}] = se[X_{(n)}] = \sqrt{\mathbb{E}[\hat{\theta}^2] - \mathbb{E}[\hat{\theta}]^2} = \sqrt{\frac{n\theta^2}{n+2} - (\frac{n\theta}{n+1})^2} = \sqrt{\frac{n\theta^2}{n+2} - \frac{(n\theta)^2}{(n+1)^2}}$$

$$\implies se[X_{(n)}] = \sqrt{\frac{(n+1)^2 \cdot n\theta^2 - (n+2)\cdot(n\theta)^2}{(n+2)\cdot(n+1)^2}} = \sqrt{\frac{n\theta^2}{(n+2)\cdot(n+1)^2}}$$

For MSE:

$$MSE[\hat{\theta}_n] = bias[\hat{\theta}_n]^2 + se[\hat{\theta}_n]^2 \quad \text{*Modeling and Inference lecture}$$

$$\implies MSE[\hat{\theta}] = MSE[X_{(n)}] = (-\frac{\theta}{n+1})^2 + (\sqrt{\frac{n\theta^2}{(n+2)\cdot(n+1)^2}})^2$$

$$\implies MSE[X_{(n)}] = \frac{\theta^2}{(n+1)^2} + \frac{n\theta^2}{(n+2)\cdot(n+1)^2} = \frac{(n+2)\theta^2 + n\theta^2}{(n+2)\cdot(n+1)^2} = \frac{2n\theta^2 + 2\theta^2}{(n+2)\cdot(n+1)^2}$$

$$\implies MSE[X_{(n)}] = \frac{2(n+1)\theta^2}{(n+2)\cdot(n+1)^2} = \frac{2\theta^2}{(n+2)(n+1)}$$

4.3 Part c

Let us determine which estimate, $2\bar{X}_n$ or $X_{(n)}$, is more efficient:

Let us compare the MSE between the two estimates:

$$\implies MSE[2\bar{X}_n] = \frac{\theta^2}{3n} \quad \text{vs.} \quad MSE[X_{(n)}] = \frac{\theta^2}{\frac{1}{2}(n+2)(n+1)}$$

So, let us compare the denominators:

$$3n \quad \text{vs.} \quad \frac{1}{2}(n+2)(n+1)$$

We can see that for $n = 1$ or $n = 2$, $MSE[2\bar{X}_n] = MSE[X_{(n)}]$. But, for $n \geq 3$, we have that $3n < \frac{1}{2}(n+2)(n+1)$ holds true. Thus, since the MSE for the estimate $X_{(n)}$ has a larger denominator for $n \geq 3$, its value will be smaller than the MSE for the estimate $2\bar{X}_n$, generally. Therefore, the estimate $X_{(n)}$ is more efficient.