Statistical Interference
Set 2

Caltech IDS/ACM/CS 157
April 2022

Dallas Taylor

1 Problem 1

In many surveys, the quantity of interest is the proportion of population units that have certain characteristic. For example, smoke or don't smoke, male or female, democrat or republican, happy with IDS 157 or not, etc. In this case, it makes sense for the unit characteristic x_i to be 1 or 0 depending on whether the characteristic is present or not, respectively:

$$x_i = \begin{cases} 1, & \text{if the } i\text{-th unit has the characteristic of interest,} \\ 0, & \text{if not.} \end{cases}$$

The population mean μ is then simply the proportion of $1s$, $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{\#1}{N}$.

Problem A: Show that the population variance in this case is

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

Solution A: *We are given a basic definition of population variance in Survey Sampling II, which we can modify in the following way:*

$$\begin{aligned}
\sigma^2 &= \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 \\
\sigma^2 &= \frac{1}{N}\sum_{i=1}^{N}(x_i^2-2\mu x_i+\mu^2) \\
\sigma^2 &= \frac{1}{N}\left[\sum_{i=1}^{N}x_i^2-2\mu\sum_{i=1}^{N}x_i+\sum_{i=1}^{N}\mu^2\right] \\
\sigma^2 &= \frac{1}{N}\left[\#1-2\mu(\#1)+N\mu^2\right] \\
\sigma^2 &= \frac{\#1}{N}-\frac{2\mu(\#1)}{N}+\mu^2 \\
\sigma^2 &= \mu-2\mu^2+\mu^2 \\
\sigma^2 &= \mu(1-\mu),
\end{aligned}$$

as desired.

Problem B: We know (lecture 4a) that s^2 is an unbiased estimate of σ^2. To compute this estimate, we need to know all sample units X_i. Use the result in (a) to derive an unbiased estimate $\tilde{s}^2$ of σ^2 that depends only on the sample mean $\bar{X}_n$ (and sample size n and population size N).

Solution B: *We are given a basic definition of s^2 in Survey Sampling II, which we can modify in the following way:*

$$\begin{aligned} s^2 &= \left(1 - \frac{1}{N}\right)\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \\ s^2 &= \left(1 - \frac{1}{N}\right)\frac{n}{n-1}(\bar{X}_n(1-\bar{X}_n)), \quad (*) \end{aligned}$$

as desired. Note that we are able to complete the step at (*) by following the same steps as part (a). The extra n in our numerator is due to the the the $\frac{1}{n}$ constant before the summation for $\bar{X}_n$.

Problem C: In a survey of a large statistics class with 300 students (say, somewhere in Hong Kong), 70 out 90 respondents said they like statistics. Using results in (b), construct a 95% confidence interval for the proportion of students in the class who like statistics.

Solution C: *We are given here that* $\alpha = 0.5$, $N = 300$, $n = 90$, $\#1 = 70$, *and thus* $\mu = \frac{70}{90} = 0.\bar{7}$. *We are given that* μ *is normally distributed. First, we want to demonstrate the value of* s, *and* $se[\bar{X}_n]$*:*

$$\begin{aligned}
s^2 &= \left(1 - \frac{1}{N}\right)\frac{n}{n-1}(\bar{X}_n(1-\bar{X}_n)) \\
s^2 &= \left(1 - \frac{1}{300}\right)\frac{90}{90-1}(0.\bar{7}(1-0.\bar{7})) \\
s^2 &= 0.174198918 \\
\therefore s &= 0.4173714389 = 0.4174
\end{aligned}$$

$$\begin{aligned}
\therefore se[\bar{X}_n] &= \frac{s}{\sqrt{n}}\sqrt{\left(1 - \frac{n-1}{N-1}\right)} \\
se[\bar{X}_n] &= \frac{0.4174}{\sqrt{90}}\sqrt{\left(1 - \frac{90-1}{300-1}\right)} \\
se[\bar{X}_n] &= 0.03687020269 = 0.0369
\end{aligned}$$

Thus, we develop our contour interval $\mathcal{I}$, *as the following:*

$$\begin{aligned}
\mathcal{I} &= \bar{X}_n \pm z_{1-\frac{\alpha}{2}} se[\bar{X}_n] \\
\mathcal{I} &= 0.\bar{7} \pm z_{1-\frac{0.05}{2}}(0.0369) \\
\mathcal{I} &= 0.\bar{7} \pm z_{0.975}(0.0369) \\
\mathcal{I} &= 0.\bar{7} \pm (1.96)(0.0369) \\
\mathcal{I} &= 0.\bar{7} \pm 0.0723 \\
\mathcal{I} &= \{0.7054, 0.8500\}
\end{aligned}$$

2 Problem 2

A survey of students at a large university is planned. The goal of the survey is to estimate the percentage of the students who own iPhones. Find the minimum sample size required to make a 95% confidence interval for the population percentage that is at most 4 percentage points wide. The population variance is unknown: assume the worst-case scenario. Ignore the finite population correction. Remark: The condition "at most 4 percentage points wide" means that the length of the interval should be at most 0.04.

Solution : *Let us use our equation for the confidence interval again. Here, we also know that we are considering our worst-case scenario for the population variance. We can determine that, since we have a quantity of interest situation similar to Problem 1, that $\sigma^2 = \mu(1-\mu)$. Thus, it is clear that our worst case variance would be where $\mu = 0.5$. Thus, we decompose our confidence interval in the following way:*

$$\begin{aligned}
\mathcal{I} &= \bar{X}_n \pm z_{1-\frac{\alpha}{2}} se[\bar{X}_n] \\
\{\bar{X}_n - 0.02, \bar{X}_n + 0.02\} &= \bar{X}_n \pm z_{0.975} se[\bar{X}_n] \\
\therefore 0.02 &= (1.96)\sqrt{\frac{\sigma^2}{n}} \\
\left(\frac{0.02}{1.96}\right)^2 &= \frac{0.5^2}{n} \\
n &= \left(\frac{0.5(1.96)}{0.02}\right)^2 \\
n &= 2401
\end{aligned}$$

Thus, we clearly find that the minimum sample size required to make a 95% confidence interval is $n = 2401$.

3 Problem 3

Solution : *See attached scripts.*

Dallas Taylor

4 Problem 4

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

Problem A: Let $\hat{\theta} = 2\bar{X}_n$. Find the bias, se, and MSE of this estimate.

Solution A: *Let us first find the bias of this estimate:*

$$\begin{aligned}
\mathit{bias}[\hat{\theta}] &= \mathit{bias}[2\bar{X}_n] \\
\mathit{bias}[\hat{\theta}] &= \mathbb{E}[2\bar{X}_n] - \theta \\
\mathit{bias}[\hat{\theta}] &= \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] - \theta \\
\mathit{bias}[\hat{\theta}] &= \frac{2}{n}\sum_{i=1}^{n}\frac{\theta}{2} - \theta \quad \textit{since we have a uniform distribution on } [0, \theta] \\
\mathit{bias}[\hat{\theta}] &= \theta - \theta = 0
\end{aligned}$$

Now, let us find the se of this estimate:

$$\begin{aligned}
\mathit{se}[\hat{\theta}] &= \mathit{se}[2\bar{X}_n] \\
\mathit{se}[\hat{\theta}] &= \sqrt{\mathbb{V}[2\bar{X}_n]} \\
\mathit{se}[\hat{\theta}] &= \sqrt{\frac{4}{n}\sum_{i=1}^{n}\mathbb{V}[X_i]} \\
\mathit{se}[\hat{\theta}] &= \sqrt{\frac{4\theta^2}{12n}} = \frac{\theta}{\sqrt{3n}}
\end{aligned}$$

Let us finally find the MSE of this estimate:

$$\begin{aligned}
\mathit{MSE}[\hat{\theta}] &= \mathit{MSE}[2\bar{X}_n] \\
\mathit{MSE}[\hat{\theta}] &= \mathit{bias}[2\bar{X}_n]^2 + \mathit{se}[2\bar{X}_n]^2 \\
\mathit{MSE}[\hat{\theta}] &= \frac{\theta^2}{3n}
\end{aligned}$$

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

Solution B: *Let us first find* $\mathbb{E}[X_{(n)}]$*:*

$$\begin{aligned}
F(x) &= \mathbb{P}[X_{(n)} \leq x] = \prod_{i=1}^{n} \mathbb{P}[X_i \leq x] = \prod_{i=1}^{n} \frac{x}{\theta} = \frac{x^n}{\theta^n} \\
\therefore f(x) &= \frac{nx^{n-1}}{\theta^n} \\
\therefore \mathbb{E}[X_{(n)}] &= \int_0^\theta x dF(x) = \int_0^\theta x f(x) dx \\
&= \int_0^\theta \frac{nx^n}{\theta^n} dx = \left[\frac{nx^{n+1}}{(n+1)\theta^n}\right]_0^\theta = \frac{n}{n+1}\theta
\end{aligned}$$

Now we can find the bias of this estimate:

$$\begin{aligned}
\textit{bias}[\hat{\theta}] &= \textit{bias}[X_{(n)}] \\
\textit{bias}[\hat{\theta}] &= \mathbb{E}[X_{(n)}] - \theta = \frac{-1}{n+1}\theta
\end{aligned}$$

Now, let us find the se of this estimate:

$$\begin{aligned}
\mathbb{V}[X_{(n)}] &= \int_0^\theta (x - \bar{X}_n)^2 dF(x) = \int_0^\theta (x - \bar{X}_n)^2 f(x) dx \\
&= \int_0^\theta (x - \frac{n}{n+1}\theta)^2 \left(\frac{nx^{n-1}}{\theta^n}\right) dx \\
&= \int_0^\theta (x^2 - 2\frac{n}{n+1}\theta x + \frac{n^2}{(n+1)^2}\theta^2) \left(\frac{nx^{n-1}}{\theta^n}\right) dx \\
&= \int_0^\theta \left(\frac{nx^{n+1}}{\theta^n} - 2\frac{n^2x^n}{(n+1)\theta^{n-1}} + \frac{n^3x^{n-1}}{(n+1)^2\theta^{n-2}}\right) dx \\
&= \left[\frac{nx^{n+2}}{(n+2)\theta^n} - 2\frac{n^2x^{n+1}}{(n+1)^2\theta^{n-1}} + \frac{n^2x^n}{(n+1)^2\theta^{n-2}}\right]_0^\theta \\
&= \frac{n\theta^2}{(n+2)} - 2\frac{n^2\theta^2}{(n+1)^2} + \frac{n^2\theta^2}{(n+1)^2} + 0 \\
&= n\theta^2 \left(\frac{1}{(n+2)} - \frac{n}{(n+1)^2}\right) \\
&= n\theta^2 \left(\frac{n^2+2n+1}{(n+1)^2(n+2)} - \frac{n^2+2n}{(n+1)^2(n+2)}\right) \\
&= \frac{n\theta^2}{(n+1)^2(n+2)}
\end{aligned}$$

$$\begin{aligned}
\therefore se[\hat{\theta}] &= \sqrt{\mathbb{V}[X_{(n)}]} \\
se[\hat{\theta}] &= \sqrt{\frac{n\theta^2}{(n+1)^2(n+2)}} = \frac{\theta\sqrt{n}}{(n+1)\sqrt{n+2}}
\end{aligned}$$

Let us finally find the MSE of this estimate:

$$\begin{aligned}
MSE[\hat{\theta}] &= MSE[X_{(n)}] \\
MSE[\hat{\theta}] &= bias[X_{(n)}]^2 + se[X_{(n)}]^2 \\
MSE[\hat{\theta}] &= \frac{1}{(n+1)^2}\theta^2 + \frac{n\theta^2}{(n+1)^2(n+2)} = \frac{(2n+2)\theta^2}{(n+1)^2(n+2)}
\end{aligned}$$

Problem C: What estimate, $2\bar{X}_n$ or $X_{(n)}$, is more efficient?

Solution C: *We can find the more efficient estimate by comparing the MSE's. Thus, it is clear that when we have the following, that $X_{(n)}$ is more efficient than $2\bar{X}_n$:*

$$\begin{aligned}
\frac{2n+2}{(n+1)^2(n+2)} &< \frac{1}{3n} \\
6n^2 + 6n &< (n+1)^2(n+2) \\
6n^2 + 6n &< n^3 + 4n^2 + 5n + 2 \\
2n^2 + n &< n^3 + 2 \\
0 &< n^3 - 2n^2 - n + 2 \\
0 &< (n^2 - 1)(n - 2)
\end{aligned}$$

The above holds true $\forall n > 2$ and thus, we have that the $X_{(n)}$ estimate is clearly more efficient.

IDS 157 PS2 - Problem 3

```
birth = readmatrix('./birth.txt');
```

Part a

Consider the birth weights stored in the variable **bwt**. First, convert the weights from ounces to kg. Compute the (population) mean weight μ. Suppose now that we want to estimate μ using simple random sampling with the sample size $n = 100$. Draw a simple random sample X from the weight population and compute the sample mean $\bar{X}_n$. Compute the exact standard error of $\bar{X}_n$. You will use this exact value as a reference value.

```
% access birth weights
bwt = birth(:,1);
bwt = bwt*0.0283495; % conversion value found online

% create sample array
n = 100;
X = bwt(randsample(length(bwt),n));

% calculate mean
X_n = 0;
for i = 1:n
   X_n = X_n + X(i);
end
X_n = X_n / n;
disp('Sample Mean:');disp(X_n);
```

```
Sample Mean:
    3.3943
```

```
% calculating standard deviation
sdev = 0;
for i = 1:n
    sdev = sdev + (X(i) - X_n)^2;
end
sdev = sqrt(sdev / n);

% calculate standard error
serr = sdev / sqrt(n);
disp('Standard Error:'); disp(serr);
```

```
Standard Error:
    0.0524
```

Part b

Assume now that X is your data and you don't know the entire population, and therefore, your can't compute the exact value of $\mathrm{se}[\bar{X}_n]$. Implement the Bootstrap method, introduced in lecture 4a. to estimate $\mathrm{se}[\bar{X}_n]$ Use $B = 10^3$ bootstrap samples. Compute the bootstrap estimate $\widehat{\mathrm{se}}_B[\bar{X}_n]$ of $\mathrm{se}[\bar{X}_n]$.

```
B = 10^3;
n = 100;
N = length(bwt);
m = 1;
serr1 = bootstrapb(X,B,N,n,m);
disp('Bootstrap Estimate of Standard Error - Part b'); disp(serr1);
```

```
Bootstrap Estimate of Standard Error - Part b
    0.0486
```

Part c

If $\frac{N}{n}$ is not an integer (where N and n are the population and sample sizes), an alternative bootstrap method was proposed in [1]. The new method differs from the original only in the way the bootstrap samples are generated, the rest is the same. In the new method, two bootstrap populations are created: the first takes k copies of each unit in the sample, and the second takes $k+1$ copies, where $N = k \times n + r, 0 < r < n$. (3) Then to sample from the bootstrap populations, first one of the two populations is chosen. The first is selected with probability $p = \left(1 - \frac{r}{n}\right)\left(1 - \frac{r}{N-1}\right)$, (4) and the second population is chosen with the remaining probability $1 - p$. Then a simple random sample of n units is taken from the chosen bootstrap population to form a bootstrap sample. Implement this modified method with $B = 10^3$ bootstrap samples and compute the bootstrap estimate $\widehat{\mathrm{se}}_{\mathrm{BF}}[\bar{X}_n]$of $\mathrm{se}[\bar{X}_n]$.

```
B = 10^3;
N = length(bwt);
n = 100;
m = 1;
serr2 = bootstrapc(X,B,N,n,m);
disp('Bootstrap Estimate of Standard Error - Part c'); disp(serr2);
```

```
Bootstrap Estimate of Standard Error - Part c
    0.0509
```

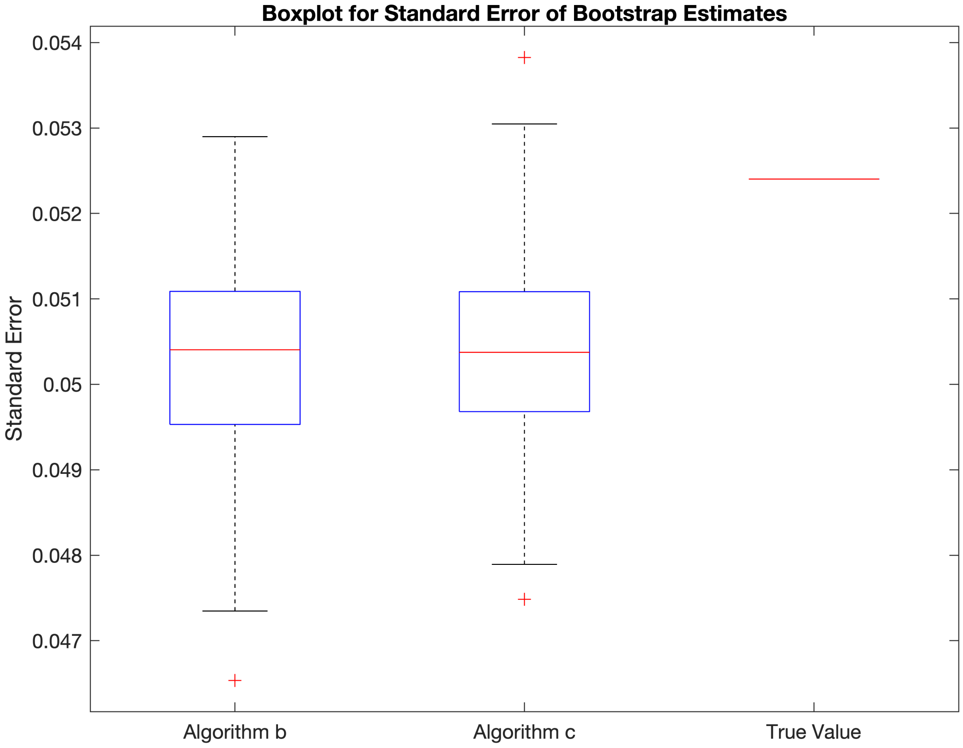
Part d

To explore the variability of the bootstrap estimates, run both algorithm in (b) and (c) 100 times, and plot two boxplots (one for each method) for the obtained estimates together with the reference true value of se[Xn].

```
B = 10^3;
N = length(bwt);
n = 100;
m = 100;
serr3 = bootstrapb(X,B,N,n,m);
serr4 = bootstrapc(X,B,N,n,m);
serr_true = repmat(serr,n,1);

figure(1); clf
```

```
g = [ones(size(serr3)); 2*ones(size(serr4));3*ones(size(serr_true))];
boxplot([serr3,serr4,serr_true],g,'Labels',{'Algorithm b', 'Algorithm c', 'True Value'
ylabel('Standard Error');
title("Boxplot for Standard Error of Bootstrap Estimates");
```



Discussion

We can see that both algorithms are approximately equally distributed. Neither distribution has the true value within its middle 50%, but both have it either close to or included in their upper range (depending on the iteration).

```
function [serr] = bootstrapb(samp,B,N,n,m)
    serr = zeros(m,1);
    rep = round(N/n);

    for l = 1:m
        % create sample arrays
        P = repmat(samp,rep,1);
        X = zeros(B,n);
        for i = 1:B
            X(i,:) = P(randsample(rep*n,n));
        end

        % calculate means
        X_n = zeros(B,1);
        for i = 1:B
```

```
            for j = 1:n
                X_n(i) = X_n(i) + X(i,j);
            end
            X_n(i) = X_n(i) / n;
        end

        % calculate mean of means
        X_n_n = 0;
        for i = 1:B
            X_n_n = X_n_n + X_n(i);
        end
        X_n_n = X_n_n / B;

        % calculate bootstrap estimate
        serrB = 0;
        for i = 1:B
            serrB = serrB + (X_n(i) - X_n_n)^2;
        end
        serr(l) = sqrt(serrB / B);
    end
end


function [serr] = bootstrapc(samp,B,N,n,m)
    serr = zeros(m,1);
    k = floor(N/n);
    r = N - n*k;
    p = (1 - r/n)*(1 - r/(N-1));

    for l = 1:m
        % create sample arrays
        P1 = repmat(samp,k,1);
        P2 = repmat(samp,k+1,1);
        curr_prob = rand(1);
        if curr_prob < p
            pop = P2;
        else
            pop = P1;
        end
        X = zeros(B,n);
        for i = 1:B
            X(i,:) = pop(randsample(length(pop),n));
        end

        % calculate means
        X_n = zeros(B,1);
        for i = 1:B
            for j = 1:n
                X_n(i) = X_n(i) + X(i,j);
            end
            X_n(i) = X_n(i) / n;
        end

        % calculate mean of means
        X_n_n = 0;
        for i = 1:B
            X_n_n = X_n_n + X_n(i);
        end
        X_n_n = X_n_n / B;

        % calculate bootstrap estimate
        serrB = 0;
        for i = 1:B
            serrB = serrB + (X_n(i) - X_n_n)^2;
        end
        serr(l) = sqrt(serrB / B);
    end
end
```