Survey Sampling
Statistics 4234/5234 — Fall 2018

Homework 2

Solutions:

1. Let $N = 6$ and $n = 3$, and consider the population

$$y_1 = 98 \quad y_2 = 102 \quad y_3 = 133 \quad y_4 = 154 \quad y_5 = 175 \quad y_6 = 190$$

for which $\bar{y}_U = 142$.

(a) For Plan 1 we have

s	Sample s	$\bar{y}_s$	$\Pr(\textit{Sample} = s)$
1	$\{1,3,5\}$	135.33	1/8
2	$\{1,3,6\}$	140.33	1/8
3	$\{1,4,5\}$	142.33	1/8
4	$\{1,4,6\}$	147.33	1/8
5	$\{2,3,5\}$	136.67	1/8
6	$\{2,3,6\}$	141.67	1/8
7	$\{2,4,5\}$	143.67	1/8
8	$\{2,4,6\}$	148.67	1/8

Thus

$$E[\bar{y}] = \sum \bar{y}_s P(s) = 135.33 \left(\frac{1}{8}\right) + \cdots + 148.67 \left(\frac{1}{8}\right) = 142$$

and

$$\begin{aligned} V[\bar{y}] &= \sum \left(\bar{y}_s - E[\bar{y}]\right)^2 P(s) \\ &= (135.33 - 142)^2 \left(\frac{1}{8}\right) + \cdots + (148.67 - 142)^2 \left(\frac{1}{8}\right) \\ &= 18.9444 \end{aligned}$$

and

$$\text{Bias}(\bar{y}) = E(\bar{y}) - \bar{y}_U = 142 - 142 = 0$$

and

$$\text{MSE}(\bar{y}) = \sum \left(\bar{y}_s - \bar{y}_U\right)^2 P(s) = V(\bar{y}) + \text{Bias}^2(\bar{y}) = 18.9444$$

For Plan 2 we have

s	Sample s	$\bar{y}_s$	$\Pr(\textit{Sample} = s)$
1	$\{1,3,5\}$	135.33	1/4
2	$\{2,3,6\}$	141.67	1/2
3	$\{1,4,6\}$	147.33	1/4

Thus

$$E[\bar{y}] = \sum \bar{y}_s P(s) = 135.33\left(\frac{1}{4}\right) + 141.67\left(\frac{1}{2}\right) + 147.33\left(\frac{1}{4}\right) = 141.50$$

and

$$\begin{aligned} V[\bar{y}] &= \sum \left(\bar{y}_s - E[\bar{y}]\right)^2 P(s) \\ &= (135.33 - 141.5)^2\left(\frac{1}{4}\right) + (141.67 - 141.5)^2\left(\frac{1}{2}\right) + (147.33 - 141.5)^2\left(\frac{1}{4}\right) \\ &= 18.0278 \end{aligned}$$

and

$$\text{Bias}(\bar{y}) = E(\bar{y}) - \bar{y}_U = 141.5 - 142 = -0.5$$

and

$$\text{MSE}(\bar{y}) = \sum \left(\bar{y}_s - \bar{y}_U\right)^2 P(s) = V(\bar{y}) + \text{Bias}^2(\bar{y}) = 18.0277 + 0.25 = 18.2778$$

(b) Plan 1 may seem more sensible, but plan 2 is actually better, since it has the lower MSE.

2. Consider a population of $N = 8$ units.

(a) For the given sampling plan the selection probabilities are

$$\pi_1 = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}$$

$$\pi_2 = \frac{1}{4} + \frac{3}{8} = \frac{5}{8}$$

$$\pi_3 = \frac{1}{8} + \frac{1}{4} = \frac{3}{8}$$

$$\pi_4 = \frac{1}{8} + \frac{3}{8} + \frac{1}{8} = \frac{5}{8}$$

$$\pi_5 = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}$$

$$\pi_6 = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}$$

$$\pi_7 = \frac{1}{4} + \frac{1}{8} + \frac{3}{8} = \frac{3}{4}$$

$$\pi_8 = \frac{1}{4} + \frac{3}{8} + \frac{1}{8} = \frac{3}{4}$$

(b) We have population

i	1	2	3	4	5	6	7	8
y_i	1	2	4	4	7	7	7	8

so $t = 40$.

Sample s	$\{y_i \in Sample\}$	$\bar{y}_s$	$\Pr(Sample = s)$
$\{1,3,5,6\}$	$\{1,4,7,7\}$	4.75	1/8
$\{2,3,7,8\}$	$\{2,4,7,8\}$	5.25	1/4
$\{1,4,6,7\}$	$\{1,4,7,7\}$	4.75	1/8
$\{2,4,7,8\}$	$\{2,4,7,8\}$	5.25	3/8
$\{4,5,6,8\}$	$\{4,7,7,8\}$	6.50	1/8

The sampling distribution of $\hat{t} = 8\bar{y}$ is

$\hat{t}$	38	42	52
Probability	1/4	5/8	1/8

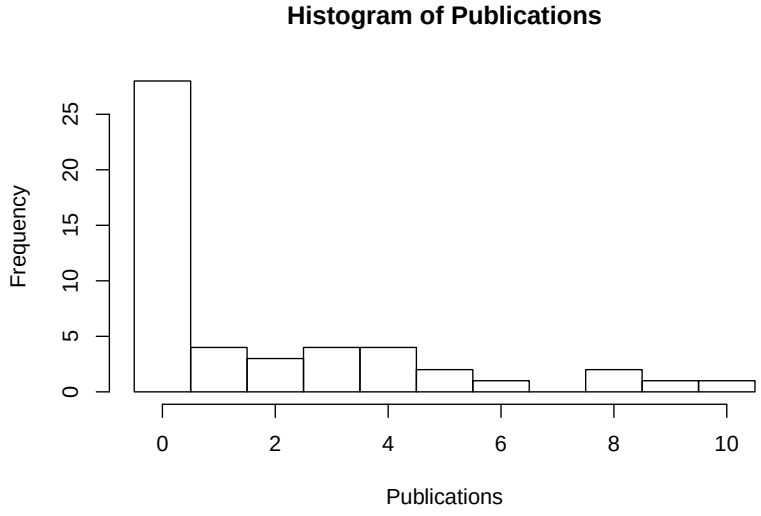
As the sampling scheme is biased toward larger values of y_i, $\hat{t}$ is a positively biased estimator of t.

3. A university has 807 faculty members. For each faculty member, the number of refereed publications was recorded. Data are available for a simple random sample of 50 faculty members.

```
> Publications <- rep(0:10, c(28,4,3,4,4,2,1,0,2,1,1))
```

(a) We can use R to construct a histogram.

```
> hist(Publications, breaks=seq(-0.5, 10.5, 1))
```



The data are right-skewed, as more than half the professors have zero publications, many have somewhere from 1 to 6, and a handful have 8 or 9 or 10.

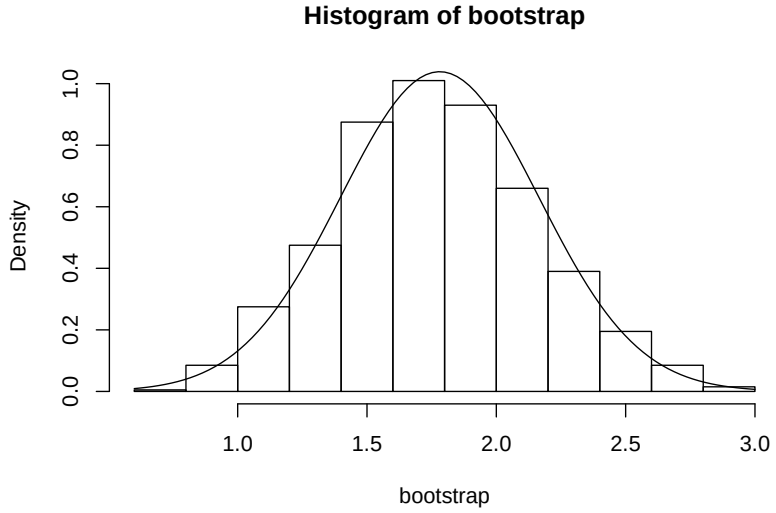
(b) Estimate the mean number of publications per faculty member by $\bar{y}$, standard error is $s/\sqrt{n} \times \sqrt{1 - n/N}$.

```
> mean(Publications)
[1] 1.78
> n <- 50; N <- 807;
> sd(Publications)/sqrt(n) * sqrt(1 - n/N)
[1] 0.3674151
```

Get $\bar{y} = 1.78$ publications, standard error of 0.37.

(c) Clearly the underlying population distribution is skewed to the right, but with sample size $n = 50$ the normal approximation to the sampling distribution of $\bar{y}$ may not be bad. Let's look at 1000 bootstrap samples.

```
> bootstrap <- rep(NA, 1000)
> for(i in 1:1000)
+ {
+  bootstrap[i] <- mean(sample(Publications, replace=T))
+ }
> hist(bootstrap, freq=F)
> curve(dnorm(x, mean(bootstrap), sd(bootstrap)), add=T)
```



Sampling distribution for $\bar{y}$ does seem to be approximately normal.

(d) Estimate the proportion of faculty members with no publications, and give a 95% confidence interval.

Let $y_i = 1$ if faculty member i had no publications, and 0 otherwise.

```
> y <- c(rep(1, 28), rep(0, 22))
> ybar <- mean(y); ybar;
[1] 0.56
```

Estimate the number of faculty members with no publications by $\bar{y} = 0.56$, standard error is $s/\sqrt{n} \times \sqrt{1 - n/N} = 0.06868$.

```
> SE <- sd(y)/sqrt(n) * sqrt(1 - n/N)
> SE
[1] 0.06868051
```

An approximate $1-\alpha$ confidence interval for $\bar{y}_U$ is $\left[\bar{y} - z_{\alpha/2}\text{SE}(\bar{y}),\ \bar{y} + z_{\alpha/2}\text{SE}(\bar{y})\right]$:

```
> ybar + c(-1,1) * qnorm(.975) * SE
[1] 0.4253887 0.6946113
```

We are 95% confident that between 42.5% and 69.5% of the faculty at this college have no refereed publications.

4. In a random sample of $n = 1000$ entries, 175 of them came from the South.

   (a) Since we are not given a population size N, we should assume it is very large relative to the sample size $n = 1000$ and thus the finite population adjustment is negligible (this is a conservative assumption).

   So we estimate the proportion of entries that come from the South by $\bar{y} = 0.175$, and the standard error is

   $$\text{SE}(\bar{y}) = \frac{s}{\sqrt{n}} = \sqrt{\frac{(.175)(.825)}{1000-1}} = 0.012$$

   A 95% confidence interval is

   $$\bar{y} \pm 1.96 \cdot \text{SE}(\bar{y}) \quad \Rightarrow \quad 0.175 \pm 1.96(0.012) \quad \Rightarrow \quad 0.175 \pm 0.0235$$

   and we are 95% confident that between 15% and 20% of entries for the last few contests came from the South.

   ```
   > y <- c(rep(1,175), rep(0,825))
   > n <- length(y); ybar <- mean(y);
   > SE <- sd(y) / sqrt(n)
   > ybar + c(-1,1) * qnorm(.975) * SE
   [1] 0.151438 0.198562
   ```

   (b) We can interpret this interval as providing *strong* evidence that the proportion of contest entries from the South is less than the proportion of persons living in the South.

5. Suppose we wish to take a simple random sample of the 580 children served by a family medical practice, to estimate the proportion who are overdue for a vaccination. If we want to estimate the proportion with 95% confidence and margin of error 0.08, we need

   $$n_0 = \left(\frac{z_{\alpha/2}S}{e}\right)^2 = \left(\frac{1.96 \times 0.50}{0.08}\right)^2 = 150$$

   and

   $$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{150}{1 + \frac{150}{580}} = 119.2$$

   so sample $n = 120$ children.