STAT 240 Homework 1

Rebecca Barter, Andrew Do and Kellie Ottoboni

February 17, 2015

Question 1. Consider a box that contains 5 "1" tickets and 7 "0" tickets. Consider drawing 6 tickets from this box at random with replacement. Let $X_1, X_2, ..., X_6$ denote the 6 numbers you observe. Let $\bar{X}$ denote the average of the draws.

a) What is $E[\bar{X}]$?

Recall that in class we showed that

$$E(\bar{X}) = \bar{t}$$

where $\bar{t}$ is the population mean. In particular, this implies that

$$E(\bar{X}) = \frac{5}{12}$$

b) What is $SE[\bar{X}]$? (R hint: Be careful whether the function "sd" divides by the square root of n or $n-1$)

Note that since this example corresponds to a simple box model with replacement, we have that

$$SE(\bar{X}) = \sqrt{Var(\bar{X})} = \sqrt{\frac{1}{n}Var(t)}$$

Using R and noting that the `sd()` function in R divides by $N-1$ rather than N, we found that (to 3dp)

$$SE(\bar{X}) = 0.201$$

c) Use R to simulate 100,000 values of $\bar{X}$. Produce a histogram of these values. (R hint: Use the function sample).

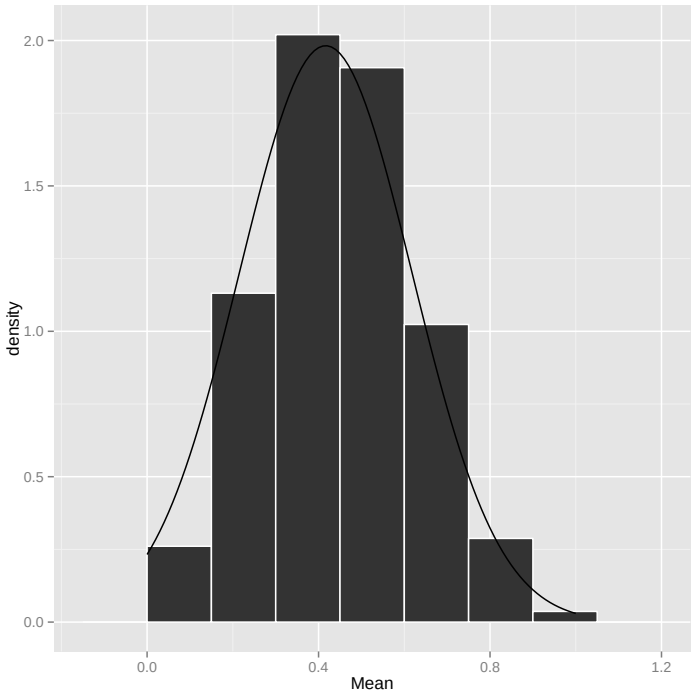

Figure 1: Histogram of 100,000 simulated values of the sample mean when the sample was taken with replacement from a box that contains 5 "1" tickets and 7 "0" tickets.

d) Let $z_1 = E[\bar{X}] + SE[\bar{X}],\ z_2 = E[\bar{X}] + 2 \times SE[\bar{X}]$, etc. For $z_1, ..., z_4$ calculate $P(\bar{X} > z_i)$ in three ways:

- **Exactly, using the binomial distribution. (Hint: It will be easier to work with the sample sum than the sample average. R hint: Use function pbinom)**
- **Estimated using the values from part (c)**
- **Using the normal approximation. Use the continuity correction. (R hint: pnorm)**

Do the same for $z_{-4}, ..., z_{-1}$ but calcualte $P(\bar{X} < z_i)$ instead of $P(\bar{X} > z_i)$. Make a table of your results and comment briefly

z	Exact	EmpiricalEst	NormalApprox
-4.00	0.00	0.00	0.00
-3.00	0.00	0.00	0.00
-2.00	0.04	0.04	0.06
-1.00	0.21	0.21	0.28
1.00	0.20	0.20	0.28
2.00	0.05	0.05	0.06
3.00	0.00	0.00	0.00
4.00	0.00	0.00	0.00

Table 1: The exact value, empirical estimation and normal approximation of the tail probability for a box that contains 5 "1" tickets and 7 "0" tickets.

We notice that the Empirical estimation using the results of our simulated value is extremely close the the exact value of the probabilities. On the other hand, the normal approximation is not nearly as accurate. This is likely because our sample size of 6 is very small and the asymptotic assumptions which underly the normal approximation are not yet accurate.

e) Repeat (a)-(d), this time sampling without replacement instead of with replacement. Use the hypergeometric distirbution instead of the binomial distribuion (R hint: phyper)

Note that since we are now sampling without replacement, we still have that

$$E(\bar{X}) = \bar{t} = \frac{5}{12}$$

but now the SE is given by

$$SE(\bar{X}) = \sqrt{Var(\bar{X})} = \sqrt{\frac{1}{n} Var(t) \left[\frac{N-n}{N-1}\right]} = 0.149$$

which is smaller than the SE when sampling with replacement.

Next, sampling 100,000 values of $\bar{X}$, we obtain the following histogram.

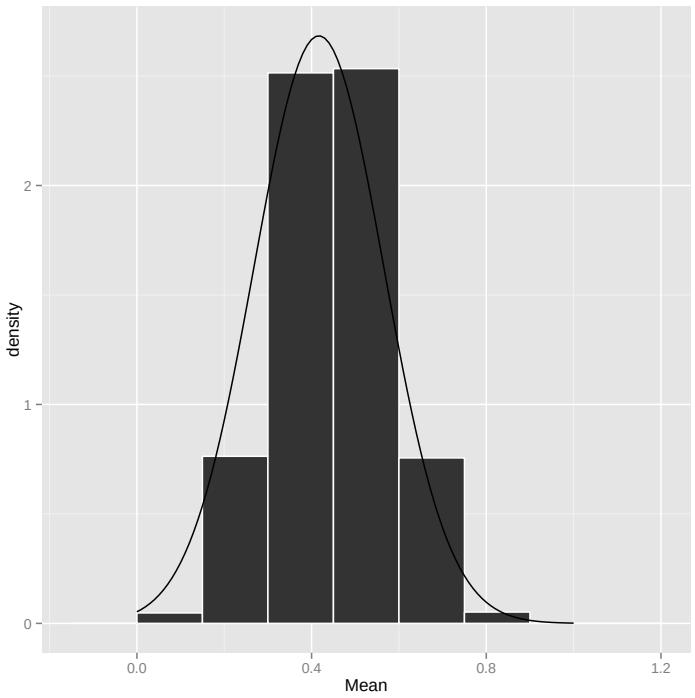

Figure 2: Histogram of 100,000 simulated values of the sample mean when the sample was taken without replacement from a box that contains 5 "1" tickets and 7 "0" tickets.

And the exact, empirical and approximated probabilities are given in the following table. Again we see that the Empirical estimation is extremely similar to the exact values, whereas the normal approximation is not nearly as accurate.

z	Exact	EmpiricalEst	NormalApprox
-4.00	0.00	0.00	0.00
-3.00	0.00	0.00	0.01
-2.00	0.01	0.01	0.07
-1.00	0.12	0.12	0.33
1.00	0.12	0.12	0.33
2.00	0.01	0.01	0.07
3.00	0.00	0.00	0.01
4.00	0.00	0.00	0.00

Table 2: The exact value, empirical estimation and normal approximation of the tail probability for a box that contains 5 "1" tickets and 7 "0" tickets.

Question 2. Repeat (1) but with a box that contains 2 "1" tickets and 98 "0" tickets.

a) - d) with replacement

Using the formulae we discussed in question 1,

$$E(\bar{X}) = \frac{2}{100}$$

$$SE(\bar{X}) = 0.057$$

Both of which are significantly smaller than the values obtained in the previous question.

Next, we plot a histogram of 100,000 simulated $\bar{X}$ values as follows:

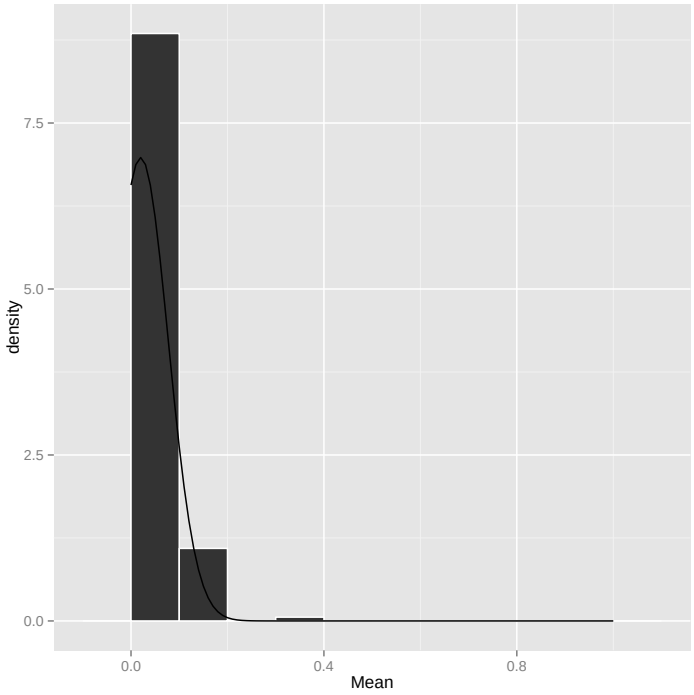

Figure 3: Histogram of 100,000 simulated values of the sample mean when the sample was taken with replacement from a box that contains 2 "1" tickets and 98 "0" tickets.

Finally, we calculate the exact, empirical and normal approximated probabilities:

z	Exact	EmpiricalEst	NormalApprox
-4.00	0.00	0.00	0.01
-3.00	0.00	0.00	0.06
-2.00	0.00	0.00	0.29
-1.00	0.00	0.00	0.68
1.00	0.11	0.12	0.68
2.00	0.11	0.12	0.29
3.00	0.01	0.01	0.06
4.00	0.01	0.01	0.01

Table 3: The exact value, empirical estimation and normal approximation of the tail probability for a box that contains 2 "1" tickets and 98 "0" tickets.

As in question 1, the empirical estimation using the results of our simulated values nearly matches the probabilities. This time, the normal approximation is particularly bad. This is occurs because we only have two "1" tickets in the box. Figure 3 shows that the empirical distribution of the sample mean is not even close to normal, as it is skewed right.

e) without replacement

Note that since we are now sampling without replacement, we have that

$$E(\bar{X}) = \frac{2}{100}$$

$$SE(\bar{X}) = 0.056$$

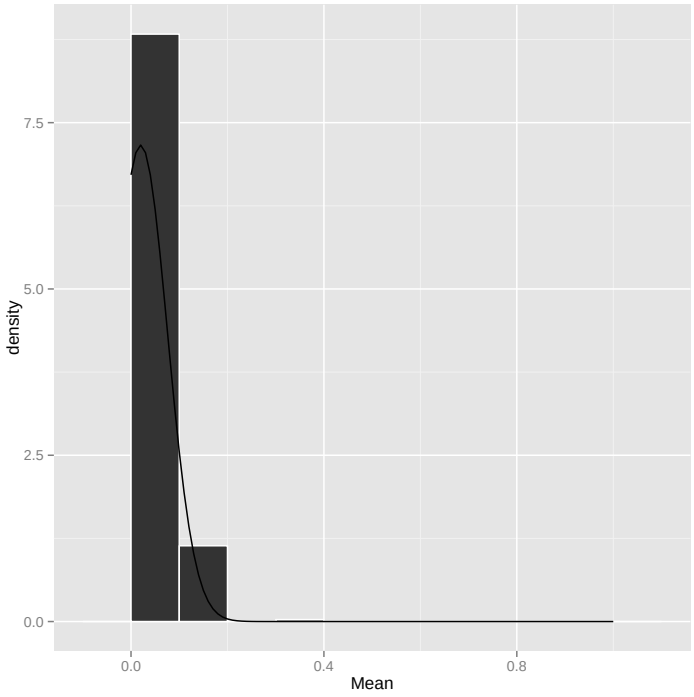

Figure 4: Histogram of 100,000 simulated values of the sample mean when the sample was taken without replacement from a box that contains 2 "1" tickets and 98 "0" tickets.

z	Exact	EmpiricalEst	NormalApprox
-4.00	0.00	0.00	0.01
-3.00	0.00	0.00	0.07
-2.00	0.00	0.00	0.31
-1.00	0.00	0.00	0.69
1.00	0.12	0.12	0.69
2.00	0.12	0.12	0.31
3.00	0.00	0.00	0.07
4.00	0.00	0.00	0.01

Table 4: The exact value, empirical estimation and normal approximation of the tail probability for a box that contains 2 “1” tickets and 98 “0” tickets.

Note that for this example, the results are extremely similar when the samples are taken both with and without replacement. Again we see that the normal approximation performs very poorly.

Question 3. Repeat (1) but with a box that contains tickets numbered "1" to "12".

a) - d) with replacement

Using the formulae we discussed in question 1,

$$E(\bar{X}) = 6.5$$

$$SE(\bar{X}) = 1.409$$

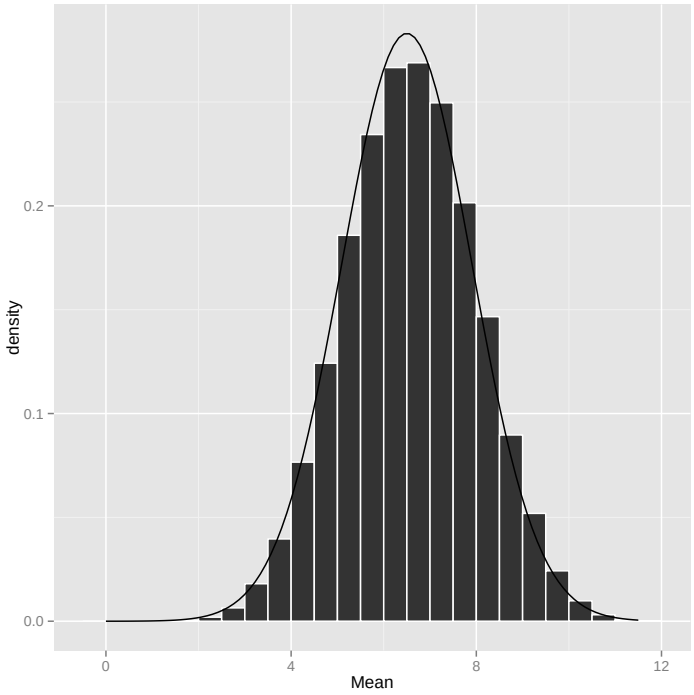

Figure 5: Histogram of 100,000 simulated values of the sample mean when the sample was taken with replacement from a box that contains tickets numbered "1" to "12".

z	EmpiricalEst	NormalApprox
-4.00	0.00	0.00
-3.00	0.00	0.00
-2.00	0.02	0.03
-1.00	0.16	0.17
1.00	0.16	0.17
2.00	0.03	0.03
3.00	0.00	0.00
4.00	0.00	0.00

Table 5: The exact value, empirical estimation and normal approximation of the tail probability for a box that contains tickets numbered "1" to "12".

In this case, the values in the box have a discrete uniform distribution. There are no "outliers", so the normal approximation performs quite well (Figure 5, Table 5), even with the small sample size of $n = 6$.

e) without replacement

Note that since we are now sampling without replacement, we have that

$$E(\bar{X}) = 6.5$$

$$SE(\bar{X}) = 1.041$$

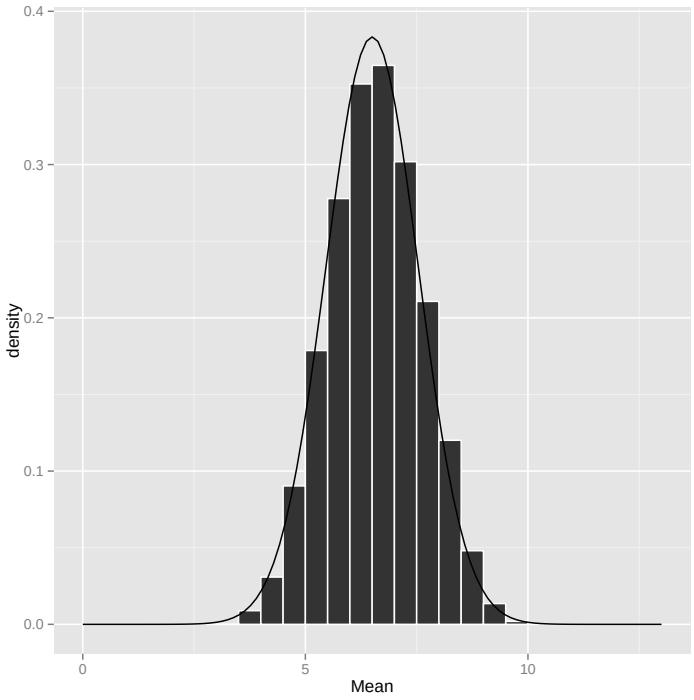

Figure 6: Histogram of 100,000 simulated values of the sample mean when the sample was taken without replacement from a box that contains tickets numbered "1" to "12".

z	EmpiricalEst	NormalApprox
-4.00	0.00	0.00
-3.00	0.00	0.00
-2.00	0.02	0.03
-1.00	0.15	0.18
1.00	0.16	0.18
2.00	0.02	0.03
3.00	0.00	0.00
4.00	0.00	0.00

Table 6: The exact value, empirical estimation and normal approximation of the tail probability for a box that contains tickets numbered "1" to "12".

When sampling without replacement in this instance, the normal approximation to the probabilities is not performing as well as when sampling with replacement, however the distribution of the $\bar{X}$'s does appear to be fairly normal.

Question 4. Repeat (1) but with a box that contains tickets numbered "1" to "11" and a ticket labelled "30".

a) - d) with replacement

Using the formulae we discussed in question 1,

$$E(\bar{X}) = 8$$

$$SE(\bar{X}) = 2.977$$

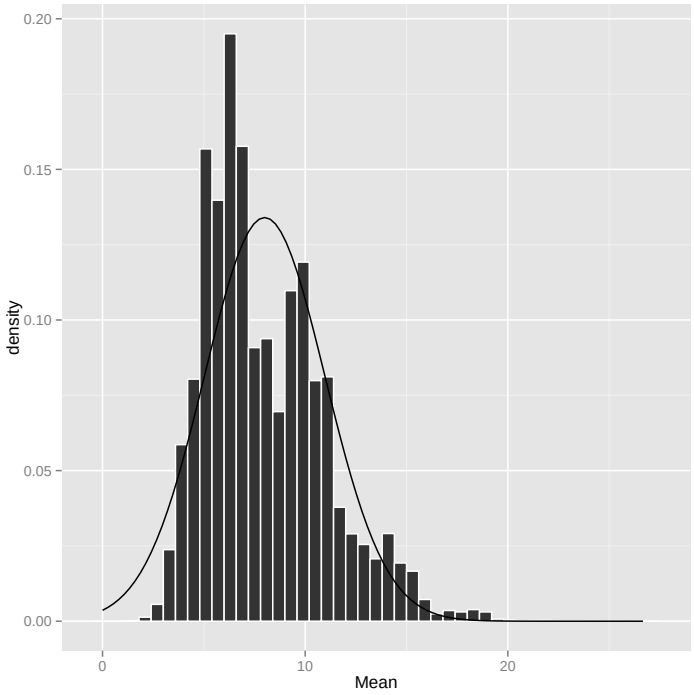

Figure 7: Histogram of 100,000 simulated values of the sample mean when the sample was taken with replacement from a box that contains tickets numbered "1" to "11" and a ticket labelled "30".

z	EmpiricalEst	NormalApprox
-4.00	0.00	0.00
-3.00	0.00	0.00
-2.00	0.00	0.02
-1.00	0.14	0.17
1.00	0.16	0.17
2.00	0.05	0.02
3.00	0.01	0.00
4.00	0.00	0.00

Table 7: The exact value, empirical estimation and normal approximation of the tail probability for a box that contains tickets numbered "1" to "11" and a ticket labelled "30".

The outlier at 30 makes the distribution of the sample mean skewed right and bimodal (Figure 7. Thus, it makes little sense to use the normal approximation. Table 7 confirms this, as the normal approximation is bad at points not near the center of the distribution.

e) without replacement

Note that since we are now sampling without replacement, we have that

$$E(\bar{X}) = 8$$

$$SE(\bar{X}) = 2.198$$

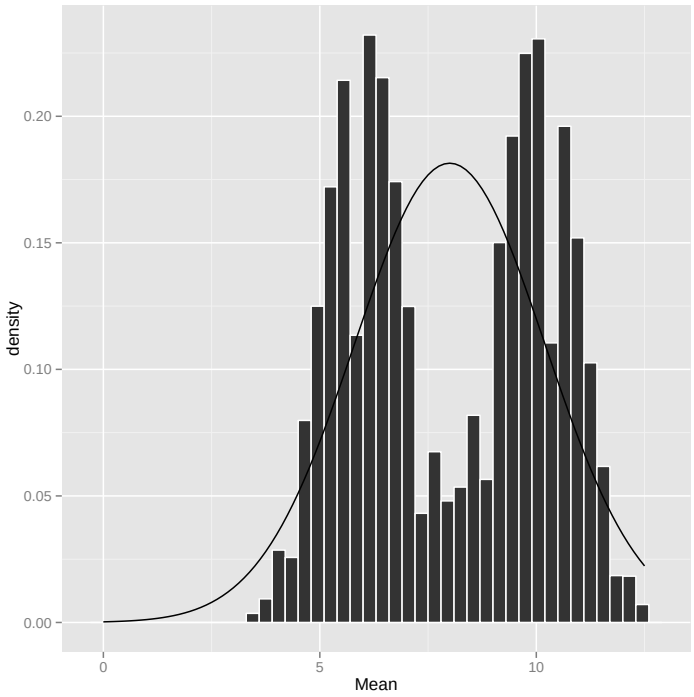

Figure 8: Histogram of 100,000 simulated values of the sample mean when the sample was taken without replacement from a box that contains tickets numbered "1" to "11" and a ticket labelled "30".

z	EmpiricalEst	NormalApprox
-4.00	0.00	0.00
-3.00	0.00	0.00
-2.00	0.00	0.02
-1.00	0.20	0.17
1.00	0.20	0.17
2.00	0.00	0.02
3.00	0.00	0.00
4.00	0.00	0.00

Table 8: The exact value, empirical estimation and normal approximation of the tail probability for a box that contains tickets numbered "1" to "11" and a ticket labelled "30".

When sampling without replacement, the distribution of the sample mean becomes bimodal and appears roughly symmetric (Figure 8). The normal approximation is a poor estimate here.

Extra Credit

Repeat for other boxes and sample sizes that you find interesting, and comment on what you learn. Along with boxes constructed "by hand" like those in (1)-(4), you may be interested to fill your box with tickets that have been randomly generated by sampling from various distributions (e.g. normal, exponential, log-normal.)

We are interested in the box in question (4), which has discrete uniform observations from 1 to 11 and a single outlier at 30. Populations like this occur in real life all the time. For example, we might be interested in the average number of dollars that somebody spent per purchase, where most purchases were small amounts but they made a few relatively large purchases.

We investigate how the normal approximation to the distribution of the sample mean improves as the sample size increases. In question (4), we sampled 6 tickets from the box. Here, we try sampling 8 and 10. When sampling with replacement, the standard errors of the sample mean with $n = 6, 8, 10$ are approximately $2.977, 2.578, 2.306$ respectively. Without replacement, the standard errors are approximately $2.198, 1.555$, and 0.983. This confirms our intuition that the sample mean becomes more concentrated around the true value of the box mean as the sample size increases. Figure 9 shows that the distribution of the sample mean is bimodal for smaller n, but smooths out and appears more Gaussian as the sample size increases.

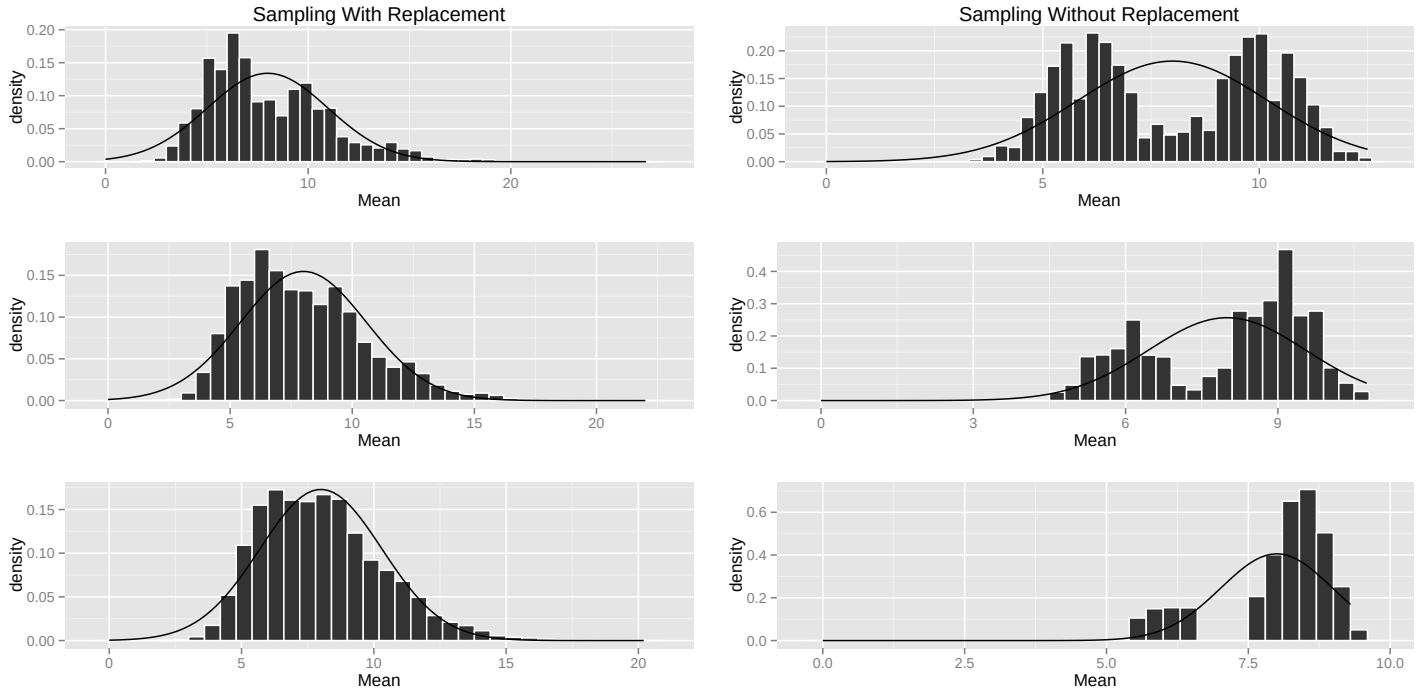

Figure 9: Distribution of the sample mean, where rows represent sampling with $n = 6, 8$, and 10.

Small Extra Credit

In class we considered the model

$$Y = a + X$$

where X is standard normal, and ask the question of whether a^2 is estimable. Is it? Justify your answer.

We claim that $Y^2 - 1$ is an unbiased estimator of a^2. The proof is as follows:

$$\begin{aligned}
E(Y^2 - 1) &= E(a^2 + 2aX + X^2 - 1) \\
&= E(a^2) + E(2aX) + E(X^2) + E(-1) \\
&= a^2 + 2aE(X) + E(X^2) - 1 \\
&= a^2 + 0 + (E(X)^2 + \mathrm{Var}(X)) - 1 \\
&= a^2 + 0 + 1 - 1 \\
&= a^2
\end{aligned}$$