STAT 240 Homework 1

Rebecca Barter, Andrew Do and Kellie Ottoboni

February 13, 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)

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

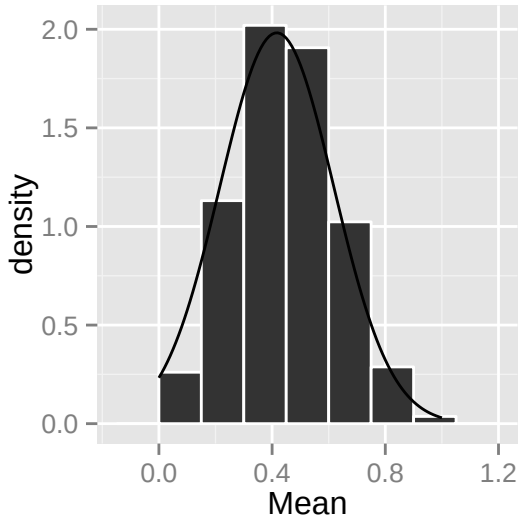

Figure 1: Histogram of 100,000 simulated values of the sample mean when the sample was taken with replacement

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

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 probability. (question 1, with replacement)

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

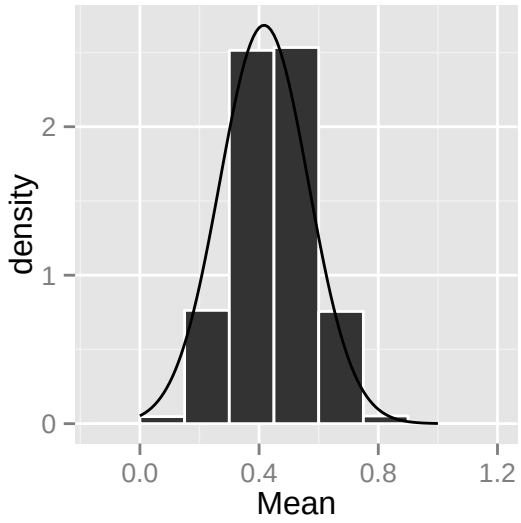

Figure 2: Histogram of 100,000 simulated values of the sample mean when the sample was taken without replacement(question 1, without replacement)

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 have that

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

and

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

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 probability. (question 1, without replacement)

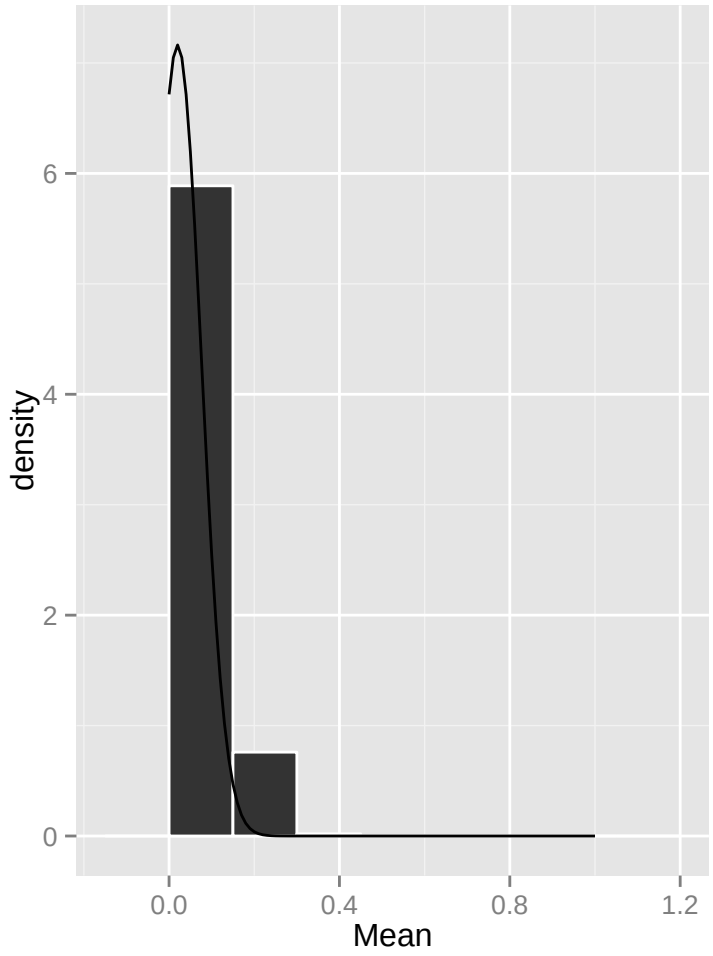

Figure 3: Histogram of 100,000 simulated values of the sample mean when the sample was taken with replacement (q2, with replacement)

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,

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

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

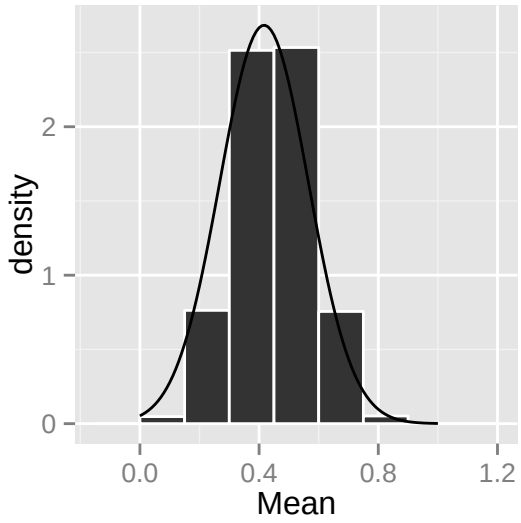

Figure 4: Histogram of 100,000 simulated values of the sample mean when the sample was taken without replacement (q2, without replacement)

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 probability. (q2, with replacement)

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

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

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

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 probability.(q2 without replacement)

The same pattern appears as in the case of sampling with replacement: the normal approximation performs very poorly.

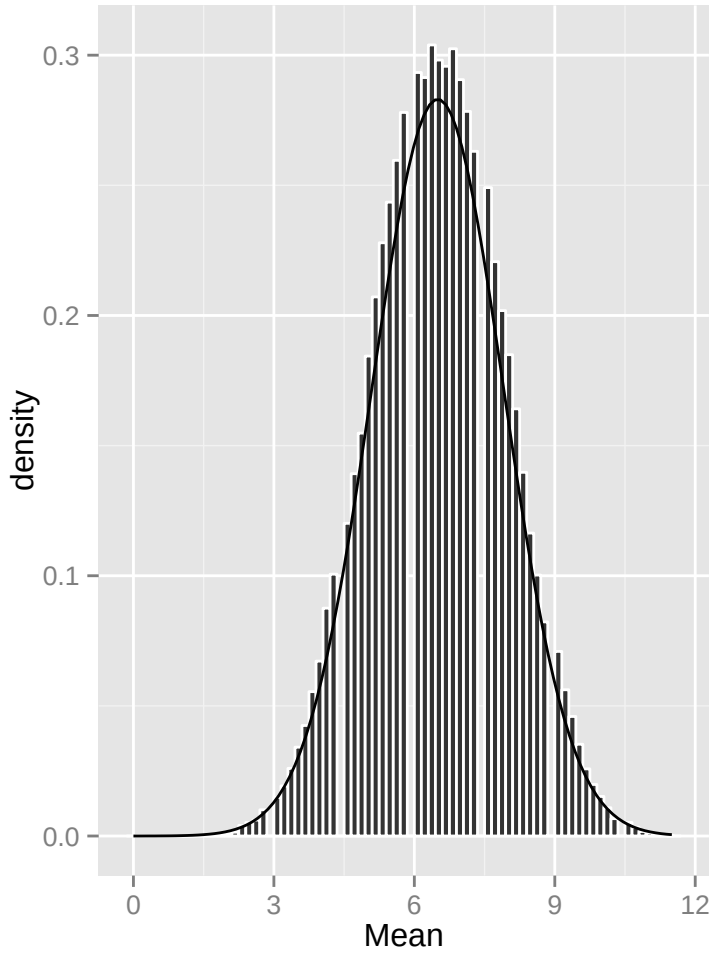

Figure 5: Histogram of 100,000 simulated values of the sample mean when the sample was taken with replacement (q3, with replacement)

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,

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

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

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 probability.(q3, with replacement)

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

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

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

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 probability.(q3, without replacement)

Similar to the previous part with replacement, the normal approximation performs quite well on uniform data sampled without replacement.

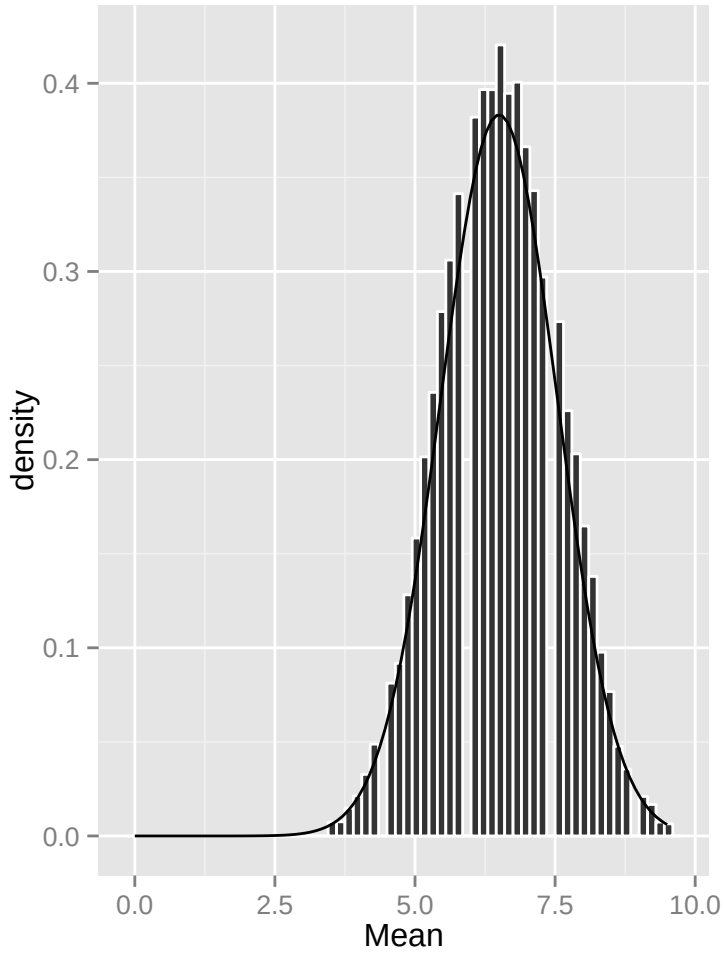

Figure 6: Histogram of 100,000 simulated values of the sample mean when the sample was taken without replacement(q3, without replacement)

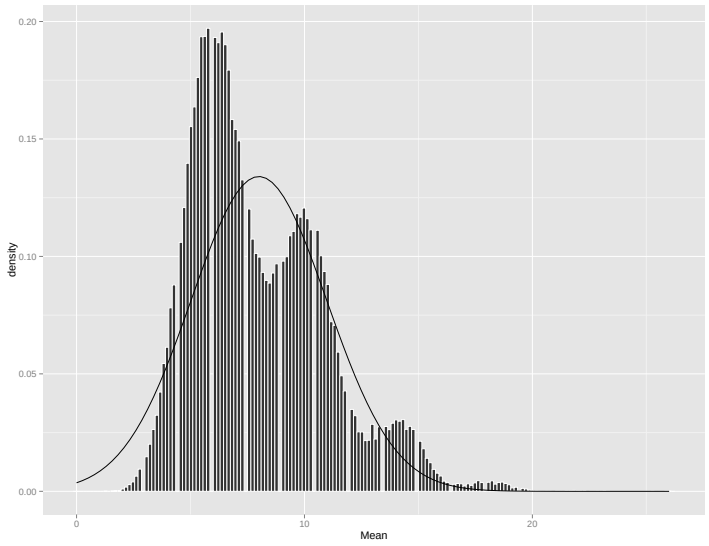

Figure 7: Histogram of 100,000 simulated values of the sample mean when the sample was taken with replacement (q4, with replacement)

Question 3. 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,

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

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

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 probability.(q4 with replacement)

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.

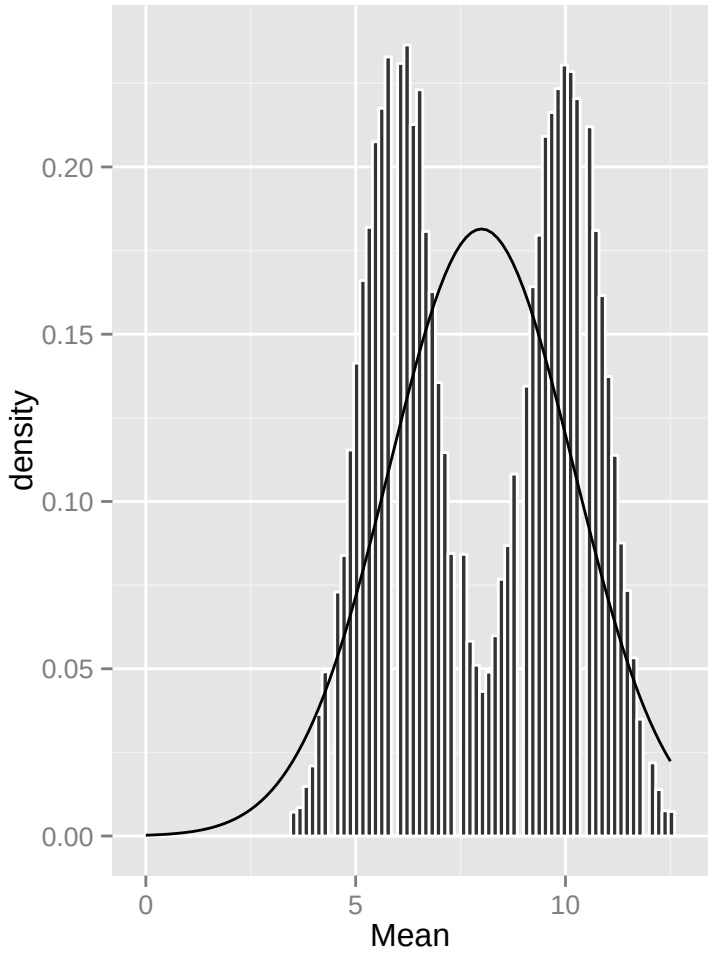

Figure 8: Histogram of 100,000 simulated values of the sample mean when the sample was taken without replacement(q4, without replacement)

e) without replacement

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

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

$$SE(\bar{X}) = 2.918$$

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 probability. (q4, without replacement)

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