Coursework: Arbitrary discrete distribution
Week 2

By Ida Johanne Austad

1. **Calculate the expected value and variance analytically and numerically:**

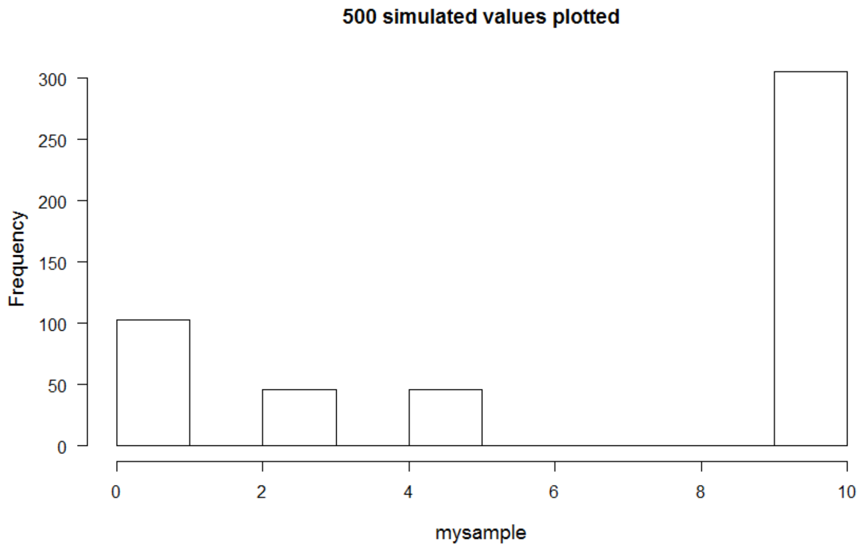
Analytical approach:

$$E(X) = \sum_{i=1}^{4} x_i \times p(x_i) = 0.2 \ \times 0 + 0.1 \times 3 + 0.1 \times 5 + 0.6 * 10 = 6.8$$

$$V(X) = E(X^2) - E(X)^2 = 0.2 \times 0^2 + 0.1 \times 3^2 + 0.1 \times 5^2 + 0.6 \times 10^2 = 0.9 + 2.5 + 60 - \ 6.8^2$$
$$= 63.4 - 46{,}24 = 17.16$$

Numerical approach: *See attached code* which gives the same answers of course.

2. **Simulate 500 values from this distribution (you can use the R function 'sample')**
   See attached code.
3. **Draw a histogram of the 500 simulated values. Compare the histogram to the specified probability distribution. Does it meet our expectation?**
   See attached code.



Yes, it does meet my expectation. The probability of getting a 10 is the largest (as given by the probability function), and a clear majority of our sample was 10's. Second runner up was 0's, with the probability of 0.2 – and we see from the histogram that we have more 0's than the two last possibilities. Lastly, our sample contains quite few 3's and 5's relative to the other two outcomes.

4. **Next, generate 500 simulated values of $\overline{x}$ where $\overline{x}$ is the mean value of 4 observations drawn from the specified discrete distribution.**
   See attached code.

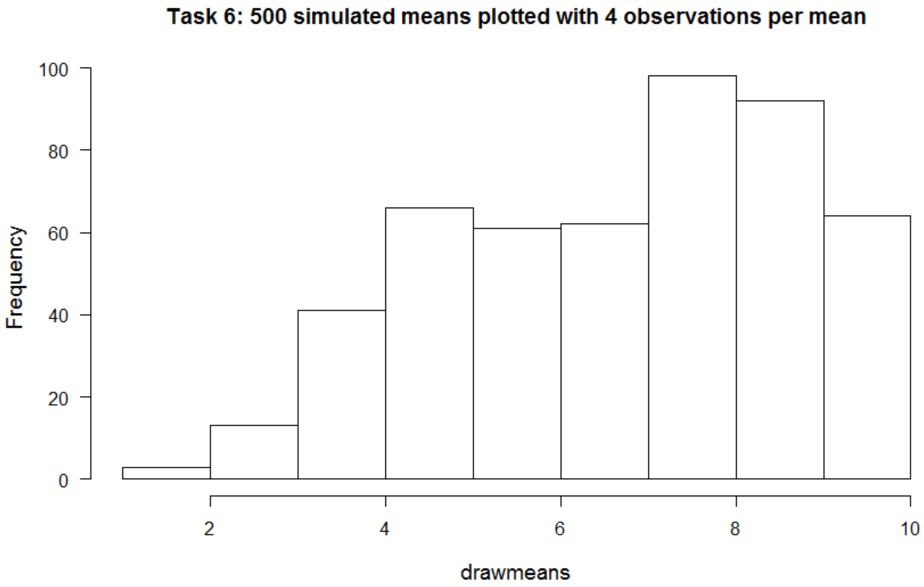
5. **Now use variance of estimator $\overline{x}$ given by $Var(\overline{x}) = \sigma^2/n$. Use R to compute this variance.**

   See attached code.

   Calculated variance of the estimator is 4.29.

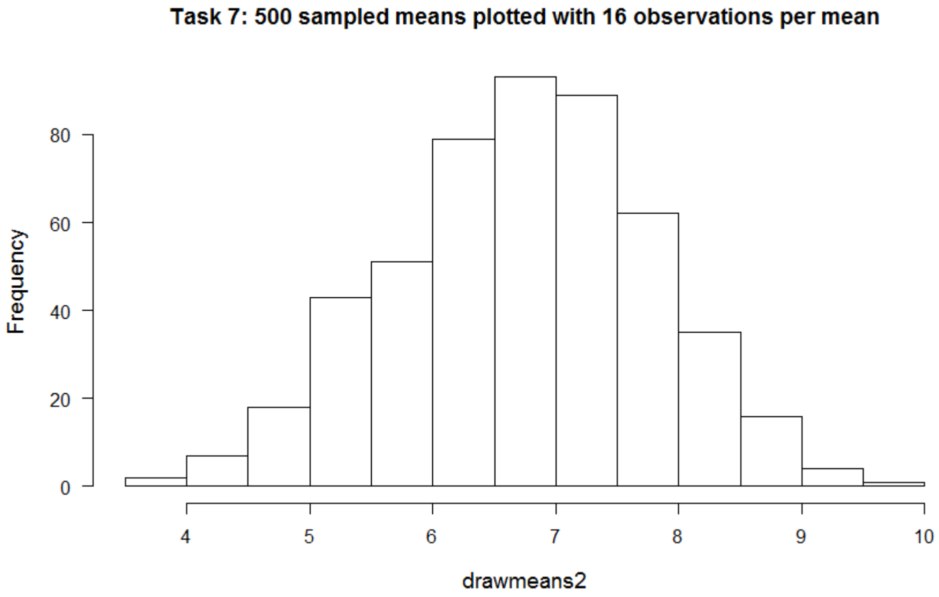
6. **Visualize the results from (4) in a histogram and interpret the results with respect to the following: What can you say about the variance of the 500 simulated means? Are they close to the theoretical value?**

   See attached code.



The theoretical value, as calculated in exercise 5, was 4.29. By looking at the plot above we can see that the spread is a bit wide and centred around the population mean (6.8 as calculated in exercise 1.). Remembering that standard deviation is the square root of the variance (should thus be a bit more than 2 in this case) we can argue that the theoretical variance is reflected in the plot of the simulated mean. As a majority of the simulated means fall within 1 standard deviation of the population mean. We can also calculate the variance of our sampled means using the var() function in R (see attached code), which for the sample above would gave a variance of 4.092441 - close to the theoretical value.

7. **Repeat step (4) and (5) but use averages over 16 observations instead of 4. Compare and comment on the results with respect to the following: unbiasedness, consistency and the central limit theorem.**
See attached code



The variance of the plotted samples above is 1.094379. The theoretical value for the same sample size is 1.0725, which is close. We see that the variance has decreased when we increased the number of observations per sample from 4 to 16, as expected.

Unbiasedness: Based on the histogram above we can see that the simulated means are distributed evenly around the theoretical value of the mean, 6.8 from exercise 1, with bell-shaped distribution where a majority of the simulated means close to the theoretical value. Thus we can argue that the estimator is unbiased.

Consistency: in task 5 -7 we have increased the number of observation per mean we calculate from 4 to 16. We also see that this has resulted in a more concentrated distribution around the theoretical mean – the variance has decreased. By testing with even larger numbers of observations per mean (I tested) we see that the concentration becomes even “stronger”. Based on this we can argue that our estimator is consistent.

Central Limit Theorem: The Central Limit Theorem (CLT) says that the distribution of sampled means will move towards a normal distribution, even though the underlying distribution is not normally distributed (for independent random variables – as we have here). And moreover, as our sample sizes increase the mean of your samples will move towards the mean of the underlying population. Based on our histograms and calculated variances with increased sized samples (4 →16), we can see that the distribution of the simulated means become more and more evenly distributed (“bell-curved”) around the theoretically calculated value of the mean. As such our examples in this coursework illustrate the CLT.

```
#code for coursework 2

x = c(0,3,5,10)
px = c(0.2, 0.1,0.1, 0.6)

#1
#Numerical estimation of expected value
expectedV <- sum(x*px)
expectedV
#Numerical estimation of variance
variance <- sum((x^2)*px) - sum(x*px)^2
variance

#2 Simulate 500 values from the given distribution
mysample <- c()
mysample <- sample(x, size = 500, replace = TRUE, prob = px)
mysample[1:20]              #just to see parts of the sample

#3 Draw histogram of 500 simulated values
par(las = 1, cex.lab = 1.2)
hist(mysample, main = "500 simulated values plotted")

#4 Generate 500 simulated values for the mean with 4 observations
per mean
drawmeans = apply(matrix(sample(x, size = 4 * 500, replace = TRUE,
prob = px), 4), 2, mean)
drawmeans                                        #to see the output

#5 Compute variance with n = 4
varianceofmean <- variance/4
varianceofmean

#6 Visualize the results from 4
par(las = 1, cex.lab = 1.2)
hist(drawmeans, main = "Task 6: 500 simulated means plotted with 4
observations per mean")
var_drawmeans <- var(drawmeans)                  #calculating the
variance of the sample
var_theoretical <- variance/4                    #theoretical
variance of the sample

#7
drawmeans2 = apply(matrix(sample(x, size = 16 * 500, replace = TRUE,
prob = px), 16), 2, mean)
drawmeans2                                      #to see the output

length(drawmeans2)
par(las = 1, cex.lab = 1.2)
hist(drawmeans2, main = "Task 7: 500 sampled means plotted with 16
observations per mean")

var_drawmeans2 <- var(drawmeans2)   #calculating the variance of the
sample
var_theoretical2 <- variance/16     #theoretical variance of the
sample
```