Sarah Kostick
Crop_Sci 545: Homework 1

Part (1) The goal of part 1 was to define a function that would generate a random variable that was a function of normally distributed random variables. The function that I chose to develop was a function of one binomally distributed and two normally distributed and variables (See R code). This function produces a random variables that follow a bimodal distribution. However, as the two means approach equality, the distribution approaches a normal distribution.

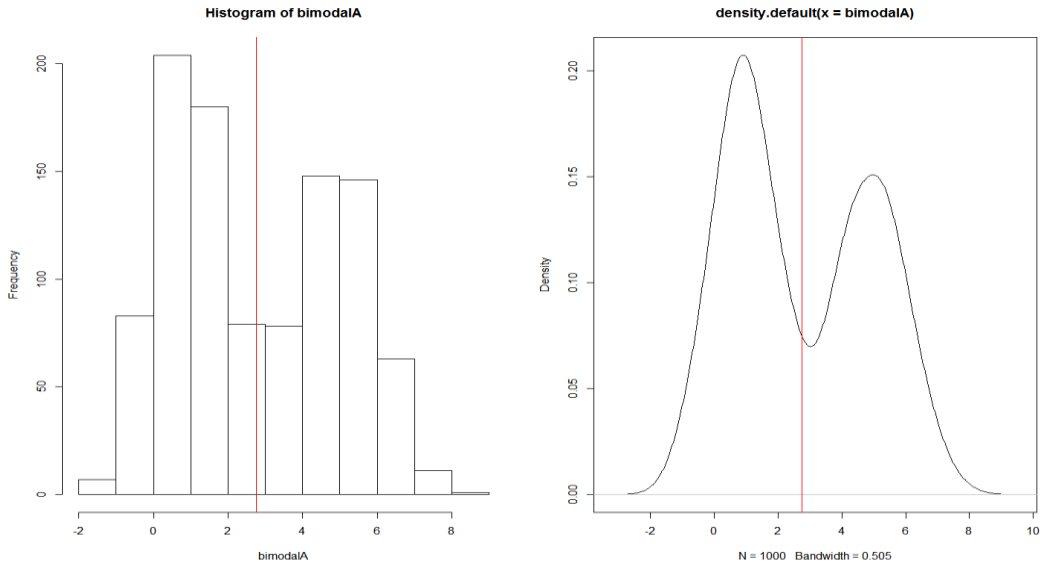
The input of this R function is

- n =number variables to be generated
- p = probability associated with the binomial portion of the function
- mu1 = mean of first normal distribution
- mu2 = mean of second normal distribution
- sig1 = standard deviation of first normal distribution
- sig2= standard deviation of second normal distribution

Example distribution

>bimodalA=rKostick(1000,0.4,mu1=1,mu2=5,sig1=1,sig2=1)

Figure 1: histogram and density plot of example bimodal distribution generated by rKostick() function in R



Part (2) The purpose of part 2 was to sample 10,000 observations from the distribution that was defined in part 1. It was expected that two modes or peaks would be observed in the density and histograms of the function as long as the means of the two peaks were different.

Methods:

1. The input of this R function was

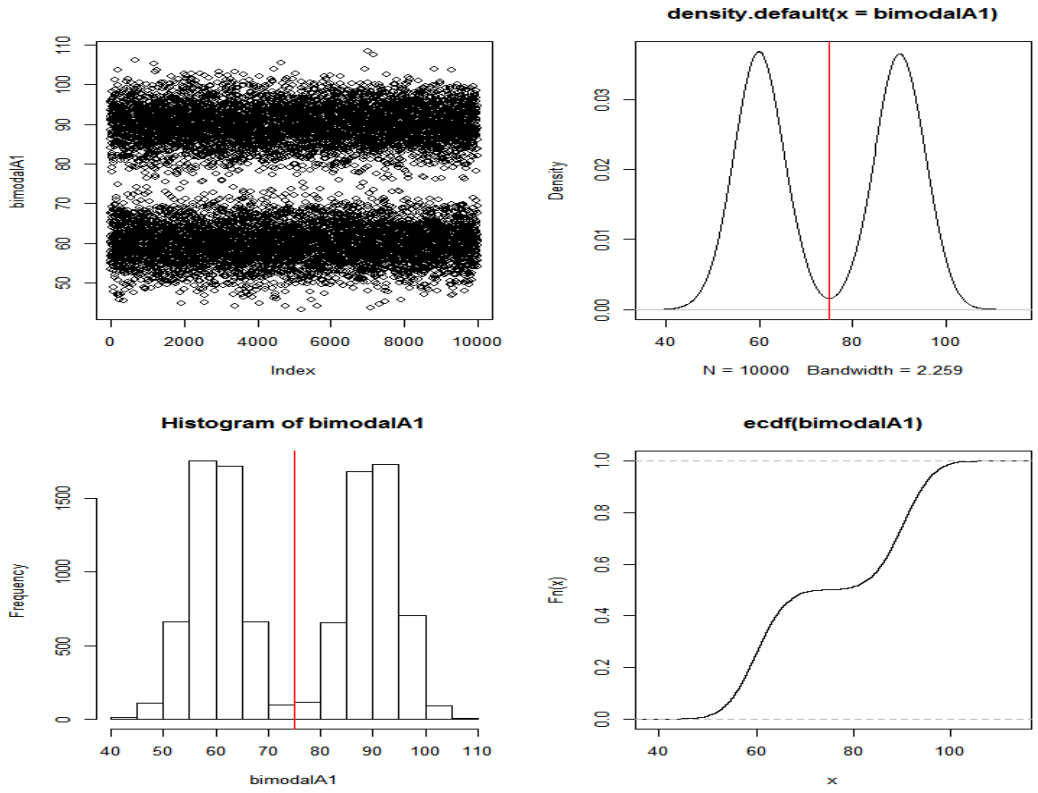
a. n = 10,000, p=0.50, mu1=60, mu2=90, sig1=5, and sig2 =5

2. Plotted scatter plot of raw data, density, histogram and cumulative probability function.
3. Calculated sample mean, variance, and range.

Results:

As expected, the majority of sample points were clustered into two main groups. These two groups, formed two main peaks in both the density plot and histogram (Figure 2). The overall mean was 74.96. The range in values was 43.35 to 108.34 and the variance of the sample was 250.87.

Figure 2: Scatter plot of raw data, density plot, histogram, and cumulative probability distribution.



Bimodal distributions are often present in nature and have use in modeling natural occurrences. For example, when studying quantitative plant diseases there are often a large number of individuals that are susceptible, an intermediate sized group that have intermediate resistance, and a large group that are resistant. This type of trait would not follow a normal distribution but a bimodal distribution. Other examples that can follow a bimodal distribution are DNA methylation in different genomes, student grades in given subjects, and peaks in vehicle traffic in the morning and evening rush hours.

Part (3) The purpose of part 3 was to define a function to generate n random variables that follow an F distribution. It was hypothesized that the function that I defined in R will give the same output as the rf() function . In order to write a function that generates random variables that follow F distribution, I followed the methods outlined below.

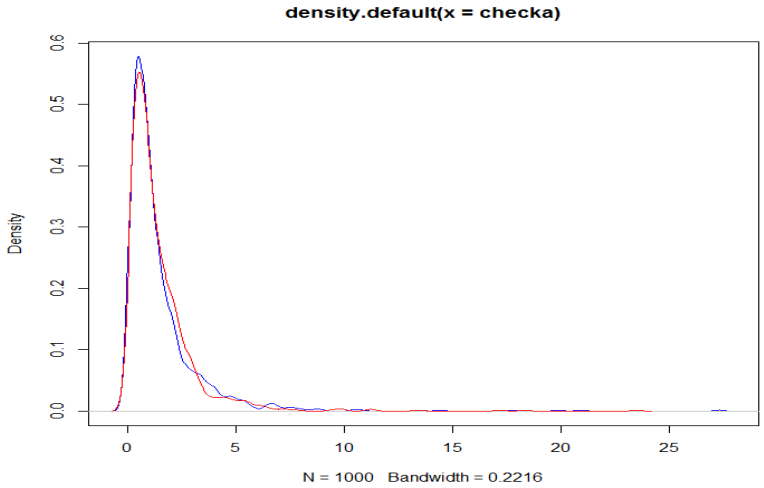
Methods:

1. Defined two normally distributed random variables (x1 and x2) using rnorm().
2. From the normally distributed variables generated random variables that follow a chi-square distribution by summing the square of the variables. These are divided by the corresponding degrees of freedom.
3. The f distribution was generated by taking the ratio of the two random variables generated in second step.

Results:

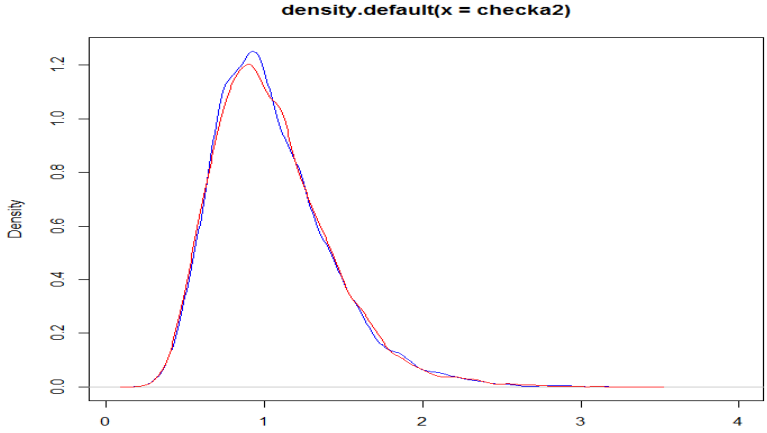
In order to check to see that my self-defined F function did generate n random variables that followed an F distribution, I generated n variables following df1 and df2 degrees of freedom using both my function and the rf() function in R. I then plotted the densities of both these samples on the same graph to compare how well they lined up. The first sample that I generated using both the rf() and my rffunct() had n=1000, df1=5, and df=6. The rf() function sample was plotted in blue and the rffunct() sample was plotted in red (Figure 3). Both density functions line up pretty well indicating that both rf() and rffunct() were generating random variables that follow an F distribution.

Figure 3: density functions for n=1000, df1=5, df2=6 variables generated by rf() and rffunct()



In order to double check, I generated a second sample using both functions that had n=10000, df1=30, df2=40. The rf() function was plotted in blue and the rffunct()in red (Figure 4). As with the first sample, both density functions lined up well indicating that both functions were generating random variables that followed an F distribution. Any variation observed is most likely due to variation in the random sampling.

Figure 4: density functions for n=10000, df1=30, df2=40 variables generated by rf() and rffunct()



Part (4) The purpose of part 4 was to sample 10, 100, 1000, and 100,000 F distributed variables and test means against the null hypothesis (H_0). In this case, variables were sampled from an F distribution with 10 and 20 degrees of freedom. The null and alternative hypothesis for this specific F distribution are listed below.

$$H_0: \text{mean} = (20)/(20-2) = 1.1111$$

$$H_A: \text{mean does not equal } 1.1111$$

Methods:

1. Sampled 10, 100, 1000, and 100,000 variables using rffunct() in R.
2. Plotted histogram of samples and calculated sample means.
3. Means were tested against H_0 using t.test() in R.

Results:

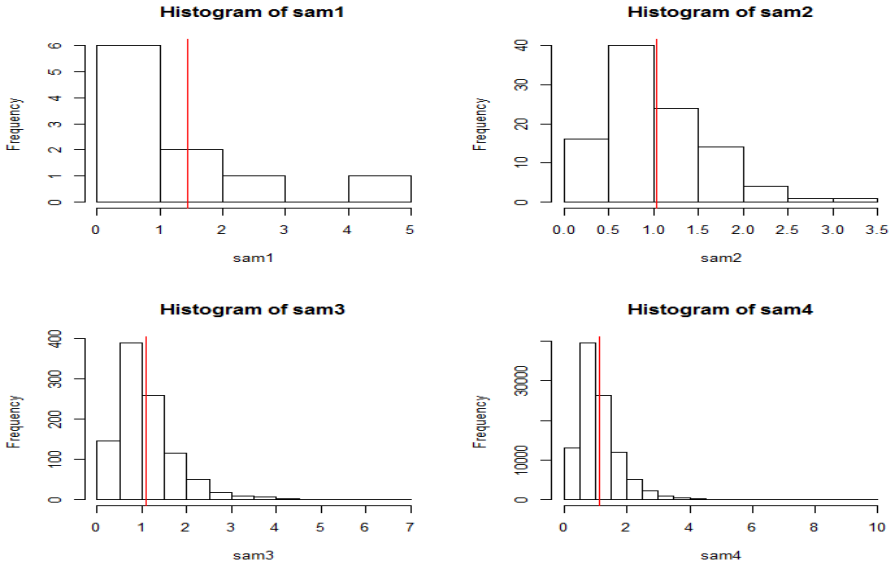
Instead of putting all the output into the report, I put the important information (e.g. mean, t value, p value) into a table format (Table 1). A sample mean was considered significantly different than the expected if the p-value was less than or equal to 0.05. It is important to note that means, t values, p values and confidence intervals will change if code is rerun and a new sample is generated.

Table 1: Sample name, sample size (n), mean, t-value, p-value and 95% confidence interval for the mean.

Sample name	Sample size (n)	Sample Mean	t-value	p value	95% confidence interval
Sam1	10	1.065	-0.438	0.672	0.829 to 1.302
Sam2	100	1.170	0.937	0.351	1.045 to 1.295
Sam3	1000	1.119	0.357	0.721	1.077 to 1.160
Sam4	100,000	1.115	1.930	0.054	1.111 to 1.119

For all samples there was no significant difference between the sample mean and the expected mean. However, the t test for sample 4 mean was 0.054, which could be considered significant. This is probably due to the effect that random sampling has on the mean. As you increase the sample size, you can detect smaller differences. As the sample size increased, the mean approached 1.1111 and the confidence interval decreased in size. This can be visualized in Figure 5. As the sample size increased, the mean approached the expected value of 1.1111 (Figure 5).

Figure 5: histogram and mean of samples 1, 2, 3, and 4.



Part (6) The purpose of part 6 was to sample 10, 100, 1000, and 100,000 F distributed variables and test sample variances against the null hypothesis (H_0). In this case, variables were sampled from an F distribution with 10 and 20 degrees of freedom. Below are listed the null and alternative hypotheses.

H_0: variance = $(2(df_2)^2(df_1+df_2-2))/(df_1(df_2-2)^2(df_2-4)) = 0.4321$

H_A: variance does not equal 1.1111

Methods:

1. Sampled 10, 100, 1000, and 100,000 variables using rffunct() in R. These samples were different than the 10, 100, 1000, and 100,000 variables that were used in part 5.
2. Plotted density of samples and calculated sample variances.
3. Variances were tested against H_0 using chi-square variance test. The p-value that was obtained was two-sided.
    a. Test: X^2 = ((n-1)*variance)/(expected variance)
    b. I also checked my answer with the sigma.test() in the TeachingDemos package.

Results:

Instead of putting all the output into the report, I put the important information (e.g. variance, X^2 value, p value) into a table format (Table 2). A sample variance was considered significantly different than the expected if the p-value was less than or equal to 0.05. It is important to note that means, X^2 values, p values and confidence intervals will change if code is rerun and a new sample is generated.

Table 2: Sample name, sample size (n), sample, F-value, p-value and 95% confidence interval for the variance.

Sample name	Sample size (n)	Sample Variance	X^2 statistic	p value	95% confidence interval
Sam1a	10	0.290	6.049	0.530	0.137 to 0.968
Sam2a	100	0.237	54.305	0.0002	0.183 to 0.320
Sam3a	1000	0.389	898.62	0.021	0.357 to 0.425
Sam4a	100,000	0.434	100460	0.299	0.430 to 0.438

The variance of two of the four samples did not differ significantly from the expected variance of 0.4321. The second and third samples that had a variance significantly different than the expected was the first sample. The significant differences between the observed and expected variances that were observed in the two samples could be simply due to random sampling. Since these are samples, we can expect some deviation from the expected. If we were to rerun these samples, it is highly possible that significant differences would not be observed. Overall, as the sample size increased, the sample variance approached the expected variance and the 95% confidence interval decreased in length.