**AML 612 Homework #2**

**For all questions, please submit your R code, and a doc file with a copy of the screen output and plots (where applicable) to** [**smtowers@asu.edu**](mailto:smtowers@asu.edu)**.**

**Due Wed February 7th at noon.**

**The code in your R files should exhibit all of the good coding practices mentioned in** [**http://sherrytowers.com/2012/12/14/good-programming-practices-in-any-language/**](http://sherrytowers.com/2012/12/14/good-programming-practices-in-any-language/)

**Include a comment header in your R file stating your name, etc**

**Please do not just copy and paste code from the examples given in class into your code.**

**Please submit your files in a format hwk2\_<last name>.R, hwk2\_<last name>.doc (this file will contain the output of your R script), hwk2\_<last name>.tex and hwk2\_<last name>.pdf**

Question 1)

As done in homework #1, find two papers on a research topic that interests you (ensure that the papers involve some sort of statistical analysis of data), and analyse the papers according to the rubric of Lacum et al. Pay careful attention to the difference between motive and objective. Make sure you include the sample sizes used in the analysis in the description of “Support” for the paper. Create a bibtex file with the annotated bibliography of the papers, and a latex file that describes the papers according to the Lacum et al rubric, with citations to the papers within the document. Make sure the references are resolved in the compiled latex! (no [?])

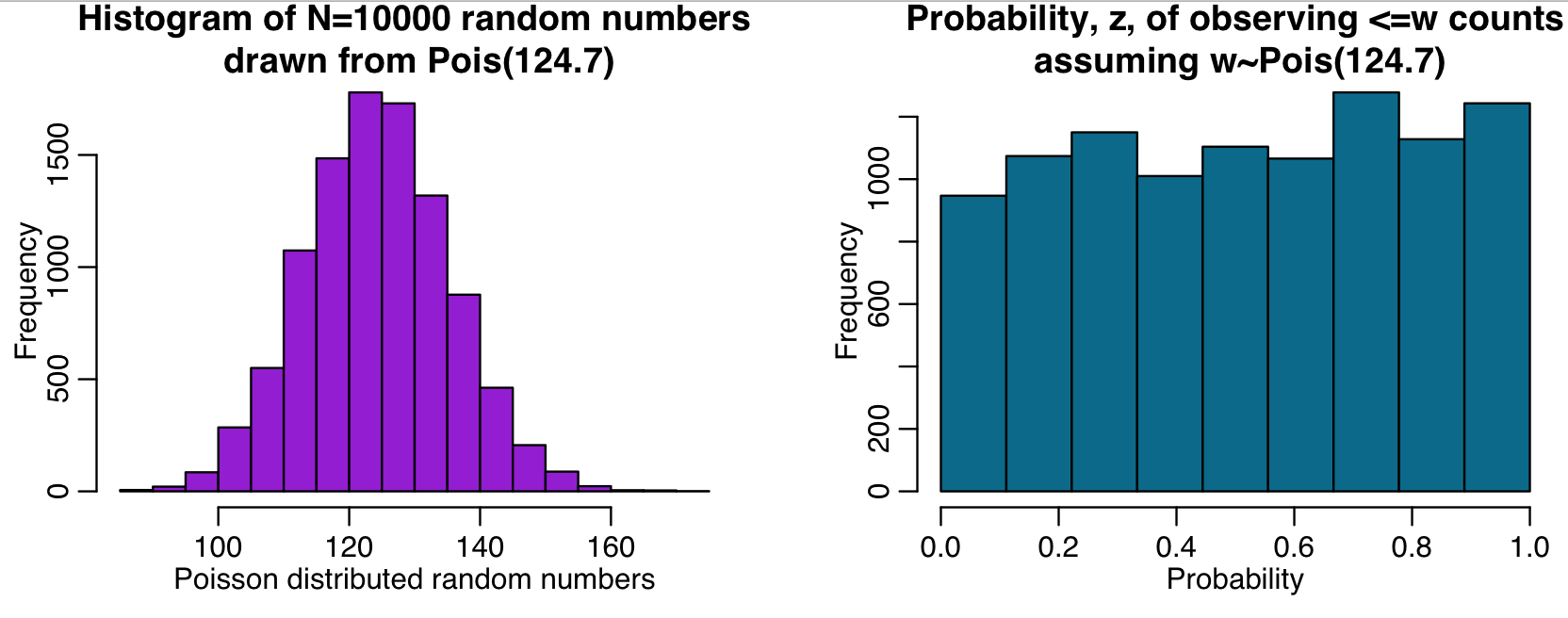
Question 2)

1. Create a vector, called x, of all odd integers between 1 to 100
2. Get the length of the vector, and print it to standard output
3. Get the mean of the vector, and print it to standard output
4. Extract every 3rd element of x (starting with the first), and put it back into x.
5. Create a vector, called y, that is the elements of x in reverse order
6. Remove the 15th element of y, and the 3rd element of x and create a new vector z that is y times x
7. Get the index of the maximum value of z, and the maximum value, and print to standard output
8. How many elements of z are divisible by 3? (note the modulus function %% comes in handy here). Print the number to standard output.

Question 3)

1. Set the random seed to be 58637. Create a vector, w, with 10,000 random numbers drawn from the Poisson distribution with mean lambda=124.7. Histogram the values, with a properly labelled plot, as below (use different colours than the ones I used).
2. Write the mean of w to standard output, along with the expected value of the mean for the Poisson distribution
3. Write the standard deviation of w to standard output, along with the expected value of the standard deviation for the Poisson distribution
4. Read the R help for cumulative probability mass function for the Poisson distribution, ppois(). Calculate a new variable z, which will be the probability of observing <=w counts, when lambda are expected with the Poisson probability distribution.
5. Read the hist() histogram function help file, focusing on the “breaks” optional argument. Histogram z into 10 bins of equal width between 0 and 1, creating the plot below, but with different colours.
6. How do you expect the histogram of z to be distributed? Does the histogram look more or less consistent with this expected distribution?
7. What fraction of the values of z are below 0.05? What fraction do you expect to be below the value of 0.05? Why?
8. Using the R function binom.test(), test to determine whether or not the fraction of values of z with z<0.05 is statistically consistent with the expected. Print the p-value of the test to standard output.

Note that the p-value returned by binom.test() tests the null hypothesis that the observed fraction is statistically consistent with the expected, vs the alternate hypothesis that the true fraction is not equal to the expected fraction. If the p-value is less than 0.05, we reject the null and accept the alternate hypothesis.



Question 4)

On the first day of class, we “randomly” sampled M&M’s out of a large bag, and counted the number of each colour of M&M’s in each of the samples. The results of this experiment are in <http://www.sherrytowers.com/m_and_m.txt>

a) What are some factors which might have potentially contributed to some samples not being a truly random sampling?

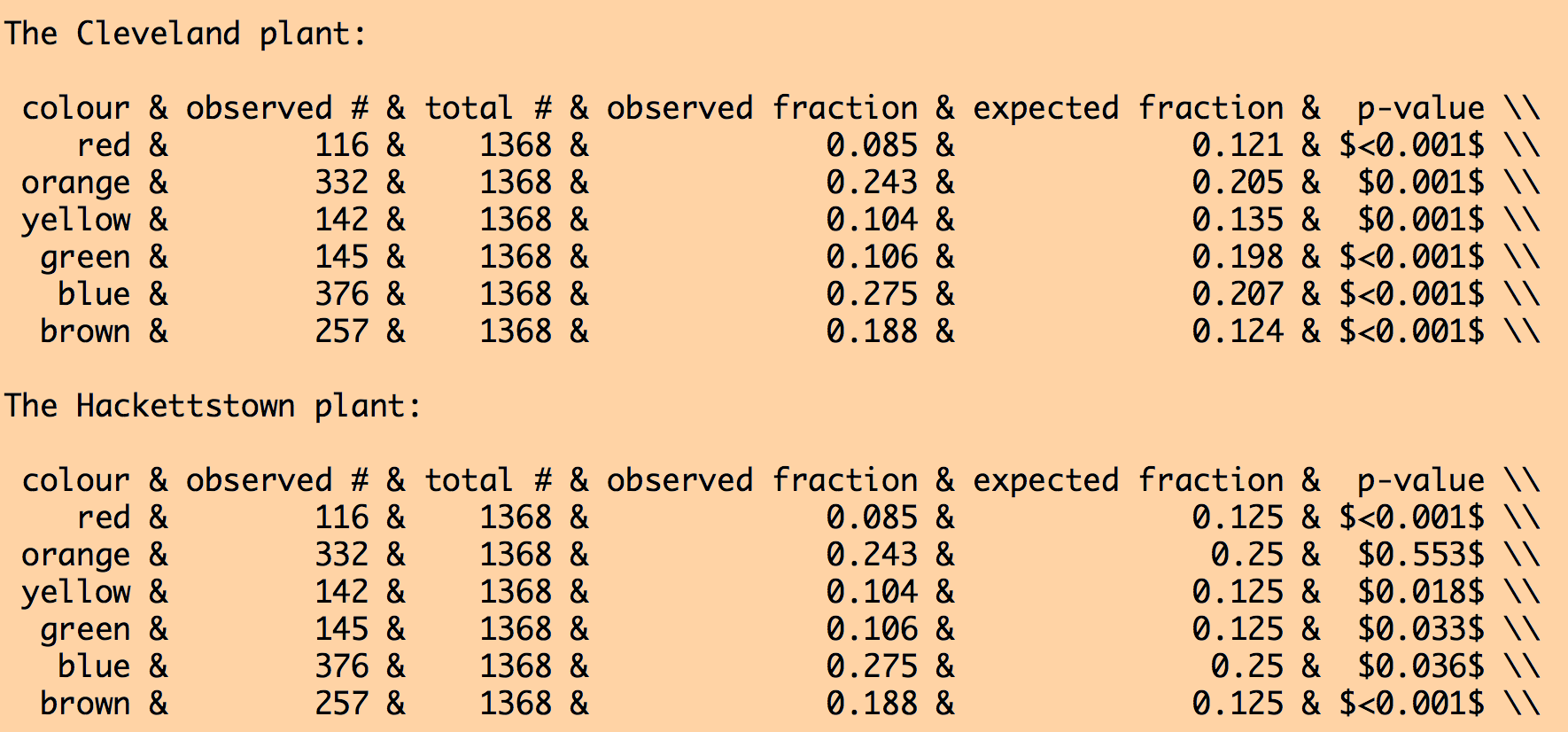
b) We’re not the first people to examine the colour distribution of M&M’s. Read, for example, this post on the topic:

<https://blogs.sas.com/content/iml/2017/02/20/proportion-of-colors-mandms.html>

It appears that the relative ratios of M&M colours have changed over time, and is even different between M&M manufacturing plants. For example, in early 2017, the above link states that the M&M company told them that the current mix of colours at the two M&M manufacturing plants in the US was:

* **CLV**: The Cleveland plant uses the following proportion of colors for plain M&M's:  
  Red=0.131, Orange=0.205, Yellow=0.135, Green=0.198, Blue=0.207, and Brown=0.124.
* **HKP**: The Hackettstown, NJ, plant uses the following proportion of colors for plain M&M's:  
  Red=0.125, Orange=0.25, Yellow=0.125, Green=0.125, Blue=0.25, and Brown=0.125.

Write an R script that calculates the estimates of the relative fractions of colours in our entire sample. Using binom.test, estimate the p-value of the observations given the expected fraction for each colour for each plant. Write the R script to format the output as below. Those of you who have used LaTex before will recognize this as a LaTex table format that is ready to be copied and pasted into a LaTex document… whenever possible, you should avoid manual transcription of analysis results into a LaTex document. Instead, always have your R scripts output the results in a format that can easily be copied and pasted into a LaTex document!



Some hints on how to get this…. I did this in less than 40 lines of code (and could have shortened it even further, in fact), and you should aim for less than 40 lines too:

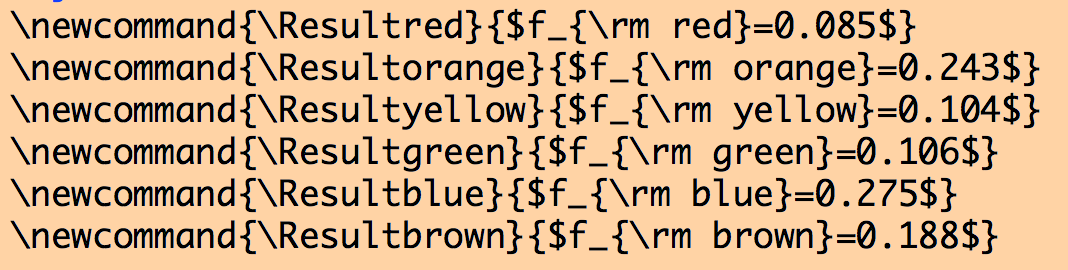
* Do a for loop, iterating twice. The first time use the expected fractions from the CLV plant, the next time through use the expected fractions from the HKP plant.
* Make a character vector of the colour names, vcolour, and a numeric vector of the expected fraction, f\_exp, for each colour.
* In the loop, do a loop over a vector of the colour names. Select the data frame column that matches that colour name. Sum the observed number for that colour in the whole sample, and the total number of M&M’s in the whole sample. Compare the fraction calculated from these numbers using binom.test to the expected fraction. Fill vectors in the loop with k=observed #, N=total #, f=observed fraction, and p=p.value from the binom.test function.
* Round the f and p vectors to three digits past the decimal point.
* For all values in the p vector that are less than 0.001, set p for those values to be the character expression “<0.001” (this will now change the p vector from numeric to character). Using the R paste() function, sandwich two “$” characters around each element of p. In LaTex, $ is the directive to enter and leave math mode. If you don’t do this, the “<” character won’t show up correctly in your compiled LaTex.
* Create and fill a data frame (called vdat, for example) with vcolour, k, N, f, f\_exp, and p. In the data.frame() function that you use to create the data frame, use the option stringsAsFactors=FALSE.
* If you were to do print(vdat) at this point, you will see a nicely formatted table. But it is missing the “&” that we need to separate columns in LaTex, and the LaTex carriage return directive “\\”.

So, create a character vector, vamp, of length the same length as vcolour, filled with “&”. Also create another vector, vend, the same length as vcolour, filled with “\\”

* Re-create the vdat data frame, this time inserting vamp between all the other variables, and the vend vector at the end. Now when you do print(vdat) you will see a nicely formatted table, suitable for copying and pasting into LaTex.

c) which manufacturing plant appears to have the colour mix most statistically consistent with our observations? Would you say that plant is the likely source of our M&M’s?

d) the \newcommand directive in LaTex allows you to define your own LaTex commands. This is particularly useful if you want to avoid manually transcribing results into a LaTex file, especially if the results are quoted multiple places in a paper. Write the R code to output the observed fractions into \newcommand format as follows (I did this in three lines of code):



Note that when using the paste() function to create this output, you’ll have to use two \\ in order to get one \ printed to the screen.

e) Write a LaTex document that has a paragraph that uses your newcommands to reference the fractions of colours in the text, and also includes the table, as shown below. Compile the LaTex document and submit both it and the compiled PDF with your homework. In the file <http://www.sherrytowers.com/AML612_Spring_2018_hwk2_example.tex> you will find a simple latex document that gives an example of defining and using \newcommands, and also gives an example of a table that is referenced in the text. Using this template to guide you, reproduce the following output in a pdf document compiled from your LaTex file:

