Binomial distribution (for lab)

Renu M. Rameshan

March 27, 2019

This document is meant to clear the confusions in lab 4 regarding binomial distribution.

The random variable under consideration is $X \sim Bin(n, p)$. In question 1 you are asked to generate different instances of X - remember what we discussed in the class, the value taken by the random variable depends on which particular outcome occurred. The function which you use in the lab in a way simulates the experiment and gives you the number of successes or an instance of X. The pmf of X is

$$f_X(k) = \binom{n}{k} p^k q^{n-k}, \quad 0 \leq k \leq n,$$

p is the probability of success in a single trial, and k is the number of successes in n trials. In the remaining part of the document I will omit the X in the subscript.

Say you have a function binornd(n,p) (this is from MATLAB). Every time you execute this function it returns a number between 0 and n (inclusive of both) which is the value taken by X for that experiment. If you use the function as binornd(n,p,1,N) it returns an array of size $1 \times N$. In this case the function does the experiment N times and gives you N instances of the random variable X. I will call this array as A. In the lab exercise N is 5000.

The values in A are always integers in the range $[0 \ n]$. If you look at the values in A you will see that certain values occur more frequently than the others and some values never occur. The value which occurs with the highest probability is obtained as follows:

$$\frac{f(k+1)}{f(k)} = \frac{n-k}{k+1}\frac{p}{q}. \tag{1}$$

Note from eq. (1) that $f(k+1)$ increases as long as

$$\frac{n-k}{k+1}\frac{p}{q} > 1$$
$$\frac{n-k}{k+1} > \frac{q}{p}$$

In the question 2 of lab you have two cases

Case 1 ($p = 0.2$, $q = 0.8$)

Here $p < q$ and the condition on k (for $f(k)$ to be increasing) is

$$k < \frac{n-4}{5} = 19.2.$$

This means $f(20)$ would be the highest value, making 20 the most frequent value for X. If you plot the pmf (which you can do using unique) you will see a peak at 20. See the blue curve in Figure 1. For $k > 20$ the probability starts decreasing.

Case 2 ($p = 0.8$, $q = 0.2$) Here $p > q$ and the condition on k (for $f(k)$ to be increasing) is

$$k < \frac{4n-1}{5} = 79.8.$$

From the above inequality, it is seen that $f(80)$ would be the peak. For $k > 80$ the pmf starts falling. This can be seen in the red plot in fig 1. For some reason MATLAB is giving 81 as the peak when I am plotting it using unique (am yet to troubleshoot it). Check what you are getting in your simulation.

Note that when p is high the values taken by the random variable is high. This is because the probability of success is high and you expect more number of successes (same as large value of k).

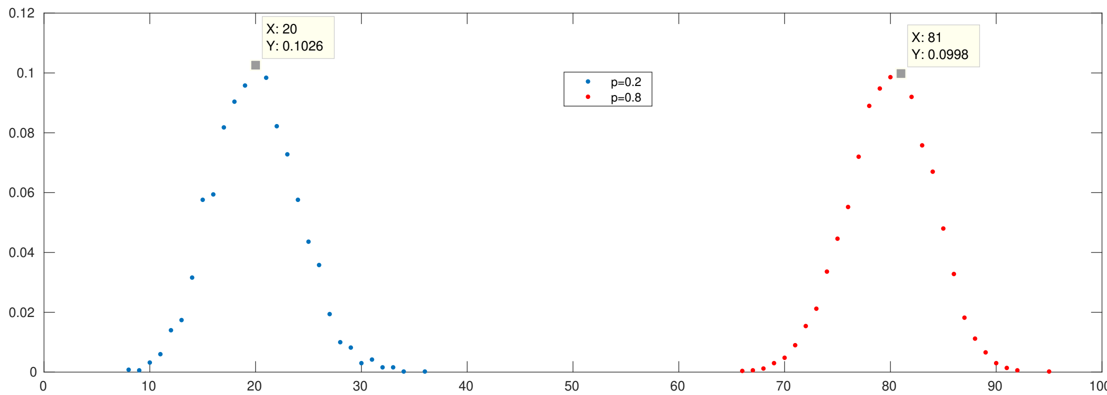

Figure 1: PMF plot for $p = 0.2$ and $p = 0.8$ from generated data.

If at all you want to plot all the instances generated (a plot with the values on Y-axis and instance number on X-axis) you get a plot as in Figure 2.

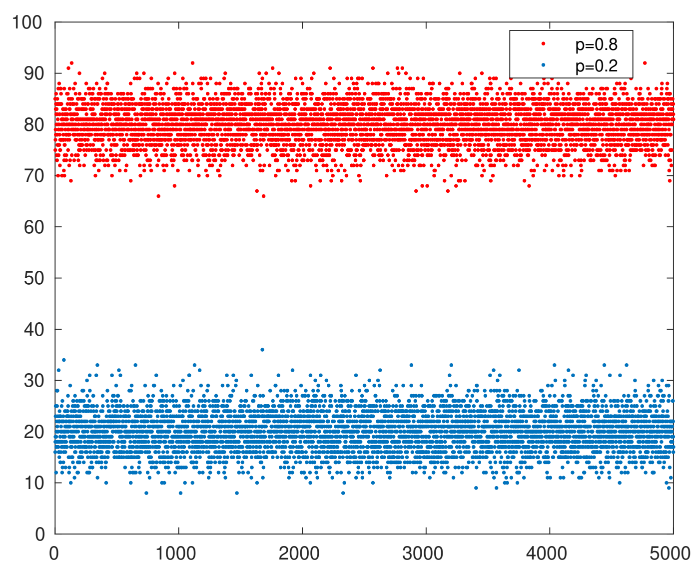

Figure 2: Plot of values taken by the RV for $p = 0.2$ and $p = 0.8$. The instance number (1 to 5000) is on the X-axis.

Note that the values coming from the distribution with higher p is larger than the one with smaller p. The reason for this was mentioned earlier (see top of this page).