Probability Homework 2

B00401062 羅文斌

Background

In this assignment, we are going to compare two distributions: **binomial distribution** and **hypergeometric distribution**. The histograms are generated with `Python` programming language. In the following sections, we will specify programming language environment, define the parameters of the two distributions, the corresponding functions in `Python`, and their differences and similarities.

Environment

All of the histograms are generated with `Python` programming language, with the following environment.

- Operating system: MacOS 10.12.3
- `Python`: 3.5.2
- `numpy`: 1.11.2
- `matplotlib`: 1.5.3
- `scipy`: 0.18.1

Python packages are imported as follows:

```python
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import binom
from scipy.stats import hypergeom
```

Parameters

To make a comparison of two distributions, we have to define the parameters used by them. Here is a table briefly summarizing the two distributions.

Distribution	**Binomial**	**Hypergeometric**
Parameters	n and p	$n1$, $n2$, and n
PMF $f(x)$	$\binom{n}{x}p^x(1-p)^{n-x}$	$\binom{n1}{x}\binom{n2}{n-x}/\binom{n1+n2}{n}$
`Python`	`scipy.stats.binom`	`scipy.stats.hypergeom`

Here, we assign statistical desctiption to the parameters based on sampling of success and failure.

- n: total number of sampling
- p: probability of success
- $n1$: total number of successes in the sampling pool
- $n2$: total number of failures in the sampling pool

`Python` Codes

For each code example, we will define the following parameters:

- n: is chosen arbitrarily.
- p: takes on 9 different values equally spaced between 0.1 and 0.9.
- $n1$ and $n2$: are chosen according to p, such that $n1 = (n1 + n2) \cdot p$, and $n2 = (n1 + n2) \cdot (1 - p)$.

After the above parameters are defined, the following code will be executed repeatedly to generate histograms.

```python
## generate subplots
figure, (binom_plt, hypergeom_plt) = plt.subplots(2, sharex = True)

## binomial distribution
binom_pmf = binom.pmf(np.arange(n + 1), n, p)
binom_title = "Binomial Distribution: n = %d, p = %.1f" % (n, p)
binom_plt.bar(np.arange(n + 1), binom_pmf, width = 1)
binom_plt.set_xlim([0, n + 1])
binom_plt.set_title(binom_title)

## hypergeometric distribution
hypergeom_pmf = hypergeom.pmf(np.arange(n + 1), n1 + n2, n1, n)
hypergeom_title = "Hypergeometric Distribution: n1 = %d, n2 = %d" % (n1, n2)
hypergeom_plt.bar(np.arange(n + 1), hypergeom_pmf, width = 1)
hypergeom_plt.set_xlim([0, n + 1])
hypergeom_plt.set_title(hypergeom_title)

## show and save figures
figure.show()
figure.savefig("p=%.1f.png" % p)
```

Differences

There are two major differences of the two distributions:

1. **Replacement**: Binomial distribution is sampling with replacement, meaning that the probability of success is independent of the number of sampling. While in hypergeometric distribution, which is sampling without replacement, the probability of success depends on how many successes and failures have been sampled.
2. **Upper bound and lower bound**: In the case of binomial distribution, there is no strict upper and lower bound on the number of successes that can be sampled. While in hypergeometric distribution, the number of sampled successes is upper-bounded by the total number of successes in the sampling pool, i.e. $n1$, and lower-bounded by the total number of successes that must be sampled. i.e. $n - n2$.

To have a better look at the effects of the above factors on the two distributions, we control $n = 10$ and $n1 + n2 = 20$, meaning that the total number of sampling is 10, and the total number of successes and failures in the sampling pool is 20. The code defining n, p, $n1$, and $n2$ is shown in Example 1, and histograms are shown in Figure 1.

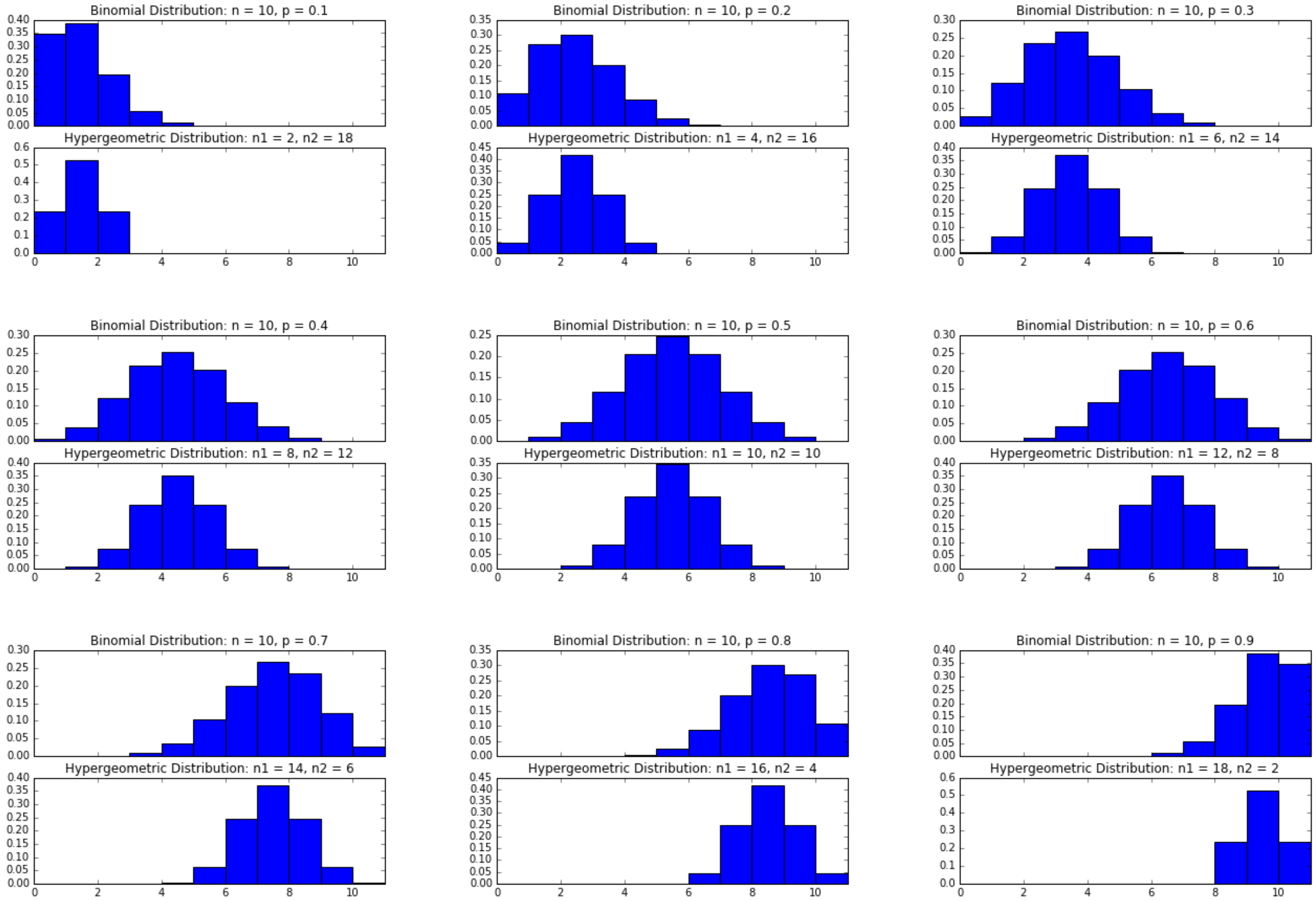
From the figure, we can see that the pmf of hypergeometric distribution is always a symmetric histogram about $n \cdot p$, while the pmf of binomial distribution shows a dissymmetric histogram whenever $p \neq 0.5$.

On the other hand, one can notice that binomial distribution is more widely-distributed, since there is no restriction on the number of successes that can be sampled. However, in the case of hypergeometric distribution, the pmf seems to be upper-bounded and lower-bounded. More precisely speaking, for any $x > n1$ and $x < n - n2$, the value of pmf must be 0.

Example 1

```
n = 10
for p in np.linspace(0.1, 0.9, num = 9):
    n1 = round(20 * p)
    n2 = round(20 * (1 - p))
```

Figure 1



Similarities

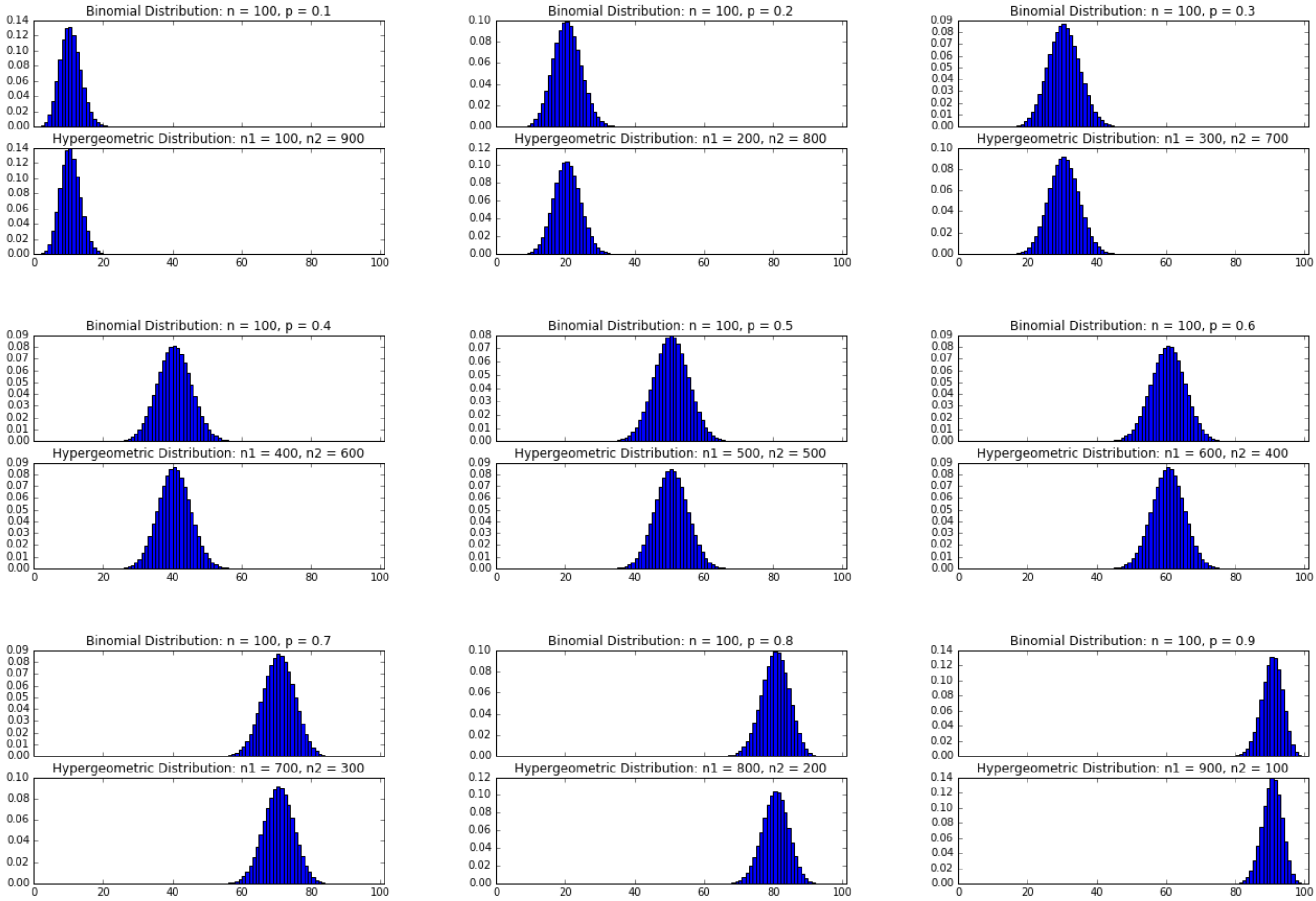
Despite the histograms of the two distributions are so different, we still can find interesting similarities between them. When $p = 0.5$, both distributions display a symmetric histogram. Also, both distributions have their pmf peak at the same value. For example, when $p = 0.2$, both distributions peak at $x = 2$. Actually, the formula of the mean of both distributions also look similar, where the mean of binominal distribution is $n \cdot p$, and the mean of hypergeometric distribution is $n \cdot \frac{n1}{n1+n2} = np$ when we let $p = \frac{n1}{n1+n2}$.

At last, we are going to show that the pmf of both distributions have virtually the same bell shape when n becomes large enough. In this example, we control $n = 100$ and $n1 + n2 = 1000$. The code defining n, p, $n1$, and $n2$ is shown in Example 2, and histograms are shown in Figure 2.

Example 2

```
n = 100
for p in np.linspace(0.1, 0.9, num = 9):
    n1 = round(1000 * p)
    n2 = round(1000 * (1 - p))
```

Figure 2


Binomial Distribution: n = 100, p = 0.1
Hypergeometric Distribution: n1 = 100, n2 = 900
Binomial Distribution: n = 100, p = 0.2
Hypergeometric Distribution: n1 = 200, n2 = 800
Binomial Distribution: n = 100, p = 0.3
Hypergeometric Distribution: n1 = 300, n2 = 700
Binomial Distribution: n = 100, p = 0.4
Hypergeometric Distribution: n1 = 400, n2 = 600
Binomial Distribution: n = 100, p = 0.5
Hypergeometric Distribution: n1 = 500, n2 = 500
Binomial Distribution: n = 100, p = 0.6
Hypergeometric Distribution: n1 = 600, n2 = 400
Binomial Distribution: n = 100, p = 0.7
Hypergeometric Distribution: n1 = 700, n2 = 300
Binomial Distribution: n = 100, p = 0.8
Hypergeometric Distribution: n1 = 800, n2 = 200
Binomial Distribution: n = 100, p = 0.9
Hypergeometric Distribution: n1 = 900, n2 = 100