

# Statistics Assignment

Vishanraj Daby  
2314620

Ihsaan Ramjaneer  
2315007

Zakariyya Kurmally  
2315839

April 20, 2024

## Contents

<b>1</b>	<b>Negative Binomial Distribution (NBD)</b>	<b>3</b>
1.1	Probability Mass Function Of NBD . . . . .	3
1.2	Expected Value and Variance of NBD . . . . .	3
1.3	Assumptions of NBD . . . . .	3
1.4	Relationship between Binomial Distribution (BD) and NBD . . . . .	4
1.5	Relationship between Geometric Distribution (GD) and NBD . . . . .	4
1.6	Illustration of NBD . . . . .	4
1.7	Issues with this application . . . . .	5
1.8	Goodness of fit test for NBD . . . . .	5
<b>2</b>	<b>Hypergeometric Distribution (HD)</b>	<b>6</b>
2.1	Probability Mass Function of HD . . . . .	6
2.2	Expected Value and Variance of HD . . . . .	6
2.3	Assumptions of a HD . . . . .	7
2.4	Relationship between HD and BD . . . . .	7
2.5	Illustration of a HD . . . . .	7
2.6	Issues with this application . . . . .	7
2.7	Goodness of Fit Test for HD . . . . .	8

# 1 Negative Binomial Distribution (NBD)

A Bernoulli trial is an experiment that can result in either a 'success' or a 'failure', but not both.

Consider a sequence of Bernoulli trials with probability of success  $p$  and probability of failure  $q$  such that  $0 \leq p \leq 1$  and  $p + q = 1$ . If  $X$  is the number of failures before the  $r^{th}$  success,  $X$  is said to follow a NBD with parameters  $r$  and  $p$ , denoted by:

$$X \sim \text{NBin}(r, p)$$

Note:

- Variance is always greater than mean for a NBD. This is known as over-dispersion
- A random variable  $D$  which follows a NBD can also be defined as the number of trials until the  $r^{th}$  success. In such a case,  $D = (X + r)$
- The terms 'success' and 'failure' in a NBD are arbitrary. As such, a NBD can also be described as modeling the number of successes before a desired number of failures. In this case, the roles of  $p$  and  $q$  are reversed

## 1.1 Probability Mass Function Of NBD

$$P(X = n) = \binom{n+r-1}{r-1} p^r q^n, \text{ for } n \in \mathbb{N}, \text{ where } q = 1 - p$$

## 1.2 Expected Value and Variance of NBD

$$E(X) = \frac{r(1-p)}{p}$$
$$Var(X) = \frac{r(1-p)}{p^2}$$

## 1.3 Assumptions of NBD

- Experiment must have 2 mutually exclusive outcomes denoted as 'success' or 'failure'
- Probability of success must be constant for each trial
- Each trial must be independent
- The experiment must have a finite number of success(es)

## 1.4 Relationship between Binomial Distribution (BD) and NBD

Consider  $n$  independent Bernoulli trials with the same probability of success  $p$ . If  $Y$  is the number of successes, it is said to follow a binomial distribution with parameters  $n$  and  $p$  denoted by:

$$Y \sim \text{Bin}(n, p)$$

Upon comparison, both the BD and NBD are based on independent Bernoulli trials. However, they differ in what they are counting. The BD counts the number of successes in a fixed number of trials  $n$  while the NBD counts the number of failures until a fixed number of successes  $r$ .

## 1.5 Relationship between Geometric Distribution (GD) and NBD

Consider a sequence of Bernoulli trials, each with the same probability of success. If  $Z$  is the number of failures before the first successful trial, then  $Z$  is said to follow a geometric distribution (GD) with parameter  $p$  representing probability of success. This is denoted by:

$$Z \sim \text{Geo}(p)$$

When comparing a GD with a NBD, it becomes apparent that the NBD generalizes a GD where instead of waiting for just the first success, we can wait for any predetermined number  $r$  of successes. In other words, a GD is a special case of NBD where  $r$  is equal to 1.

## 1.6 Illustration of NBD

To illustrate, we will use data from MES pertaining to grades of student in Economics A Level during the 2023 seating. Below is a summary of the data collected:

Grade	Point Range	$f_i$
A*	129-180	75
A	113-129	261
B	95-112	435
C	83-95	419
D	71-83	513
E	60-71	490
F	0-60	495

From the above data, we can calculate the probability that a random student has obtained a credit:

$$P(\text{credit}) = \frac{1703}{2688} = 0.6336$$

Note: a credit is awarded to students who have gotten at least a D.

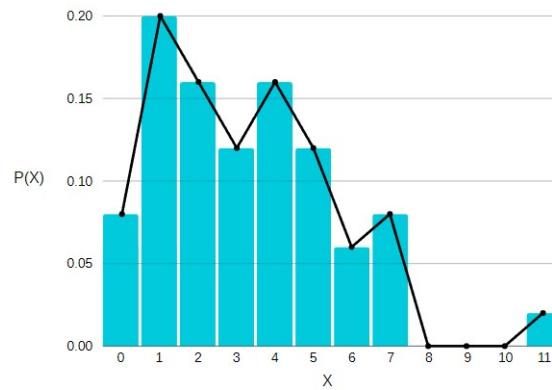
Let 'success' be a student who was awarded a credit in A-Level Economics, and 'failure' be a student who was not able to get a credit.

Our goal is to calculate the probability of  $X$  which is a random variable representing the number of failures until we obtain 5 success.

Note: probability that a student obtains a credit is 0.6336.

Using the `rbinom(50, 5, 0.6336)` function in R we have generated the following data for our distribution:

x	0	1	2	3	4	5	6	7	8	9	10	11
f	4	10	8	6	8	6	3	4	0	0	0	1
P(X=x)	0.08	0.20	0.16	0.12	0.16	0.12	0.06	0.08	0	0	0	0.02



## 1.7 Issues with this application

Since we have used the `rhyper` function in R to generate our distribution data, it may not reflect actual real life data.

## 1.8 Goodness of fit test for NBD

Based on our distribution data, the mean and variance can be calculated:

$$E(x) = \frac{\sum x \cdot f}{\sum f} = \frac{163}{50} = 3.26$$

$$Var(X) = \frac{(x_i - \mu)^2}{n - 1} = 5.46$$

Since  $\text{mean} < \text{variance}$ , the data does satisfy the over-dispersion assumption of a NBD.

The Chi-Square test will be used to determine if our distribution follows a NBD allowing a significance level of 5%.

$x$	$O_i$	$E_i$ (4 d.p)	$(O_i - E_i)^2/E_i$ (4 d.p)
0	4	5.106	0.2394
1	10	9.3534	0.04476
2	8	10.2813	0.5062
3	6	8.7898	0.8855
4	8	6.4412	0.3772
5	6	4.2481	0.4124
6	3	2.5942	0.0635
7	4	1.4936	0.2060
8	0	0.8209	0.8209
9	0	0.4345	0.4345
10	0	0.2229	0.2290
11	1	0.1113	0.0689
	50	49.8968	15.2882

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 15.2882$$

$$x_{11}^2(0.05) = 19.675$$

Since  $x^2 < x_{11}^2(0.05)$ , assuming a 5% significance level, there is statistical evidence that the random variable  $X$  does in fact follow a negative binomial distribution.

## 2 Hypergeometric Distribution (HD)

Consider a population of  $N$  objects which are divided into 2 types: type A and type B. There are  $n$  objects of type A and  $N - n$  objects of type B. Suppose a random sample of size  $r$  is taken (without replacement) from the entire population of  $N$  objects. If  $X$  is the number of objects of type A in the sample, then  $X$  follows a HD with parameters  $n$ ,  $N - n$  and  $r$  denoted by:

$$X \sim \text{HGeom}(n, N - n, r)$$

### 2.1 Probability Mass Function of HD

$$P(X = k) = \frac{{}^nC_k \cdot {}^{(N-n)}C_{(r-k)}}{{}^NC_r}, \text{ for } \max\{0, r - (N - n)\} \leq k \leq \min\{r, n\}$$

### 2.2 Expected Value and Variance of HD

$$E(X) = \frac{nr}{N}$$

$$\text{Var}(X) = \frac{nr}{N} \cdot \frac{N - r}{N} \cdot \frac{N - n}{N - 1}$$

## 2.3 Assumptions of a HD

- Finite population
- Population can be separated into 2 types
- Sampling is done without replacement (dependent trials).

## 2.4 Relationship between HD and BD

A BD is not suitable for the random variable  $X$  as the Bernoulli trials (where a “success” is represented by Type A) are dependent (since sampling is done without replacement). However if the sample size is small enough ( $< 5\%$  of population size), the BD can be used to provide a reasonable approximation for a HD.

## 2.5 Illustration of a HD

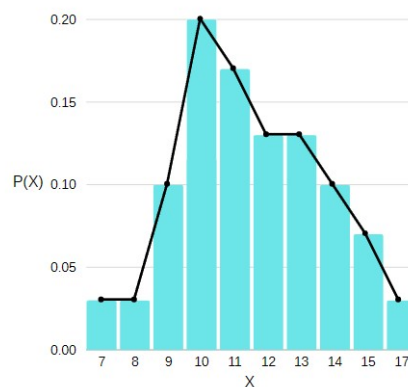
Here, we will use data from MES concerning the number of students who sat for the A-Level exams in 2023 in Rodrigues. The following table shows amount of students classified by gender:

Male	Female	Total
94	150	244

**Scenario:** Let a sample of 30 students be taken from the population without replacement. Our goal is to calculate the probability of  $X$ , which represents the number of males in the sample

The data for our distribution was generated using the `rhyper(30, 94, 150, 30)` function in R:

x	7	8	9	10	11	12	13	14	15	17
f	1	1	3	6	5	4	4	3	2	1
P(X=x)	0.03	0.03	0.10	0.20	0.17	0.13	0.13	0.10	0.07	0.03



## 2.6 Issues with this application

This particular distribution as well may not be suitable to model other applications since it only concerns itself with Rodrigues year 2023.

## 2.7 Goodness of Fit Test for HD

The Chi-Square test will be used to test if a HD is suitable for our distribution, allowing a 5% level of significance.

$x$	$O_i$	$E_i$ (4 d.p)	$(O_i - E_i)^2/E_i$
7	1	0.9165	0.0076
8	1	1.7910	0.3493
9	3	2.9187	0.0023
10	6	4.0077	0.9904
11	5	4.6722	0.0230
12	4	4.6515	0.0913
13	4	3.9711	0.0002
14	3	2.9148	0.0025
15	2	1.8423	0.0135
17	1	0.4704	0.5963
	30	28.1562	2.0764

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 2.0764$$

$$x_9^2(0.05) = 16.919$$

Since  $x^2 < x_9^2(0.05)$ , assuming a 5% significance level, there is statistical evidence that the random variable X does follow a hypergeometric distribution.