3G Hypergeometric

Contents

3G Hypergeometric

[ToC]

G.1 HG

$X \sim HG(n, m, N)$ **Analogy:** N balls in an urn, of which m are red. Pick n at once. X= number of red balls.

$$\text{pmf}: \quad p(x) = p(X = x) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}} \qquad **$$

$$\text{CDF}: \quad F(x) = P(X \le x) = \sum_{k=0}^{x} p(x)$$

$$\text{mean}: \quad E(X) = np \qquad\qquad \text{var}: \quad V(X) = \left(\frac{N-n}{N-1}\right) np(1-p)$$

$$\text{MGF}: \quad M(t) = DoesNotExist$$

where $p = m/N$.
** for $\max(0, n - N + m) \le x \le \min(n, m)$, and 0 otherwise.

```
dhyper(2,  m, N-m, n)        # pmf at x=2  (In R,  #Red, #White, #To Pick)
phyper(2,  m, N-m, n)        # CDF at x=2
phyper(.5, m, N-m, n)       # Inv CDF at q=.5
rhyper(1000, m, N-m, n)     # random sample of size 1000
```

G.2 Derivation of Hypergeometric pmf

G.3 Binomial vs Hypergeometric

- Suppose you have populatio of N subject, of which m are defective. Let

$$X = \text{[number of defectives in sample]} \ .$$

- Sample with replacement.

- Sample without replacement.

G.4 R code for Hypergeometric(n, m, N)

x =[number of heads]
n =[number of balls picked]
m =[number of Red balls]
N =[number of total balls]

```
dhyper(3,4,6,5)        #- p(3):  pmf of Hypergeoretric(m=4, N-m=6 ,n=5) at x=3   (That means N=10)
phyper(3,4,6,5)        #- F(3):  CDF of Hypergeoretric(m=4, N-m=6 ,n=5) at x=3


layout( matrix(1:2, 1, 2) )  #- Make plot layout side by side

x <- 0:10
plot(x, dhyper(x, 4,6,5), type="h", ylim=c(0,1))   #- PMF plot -
plot(x, phyper(x, 4,6,5), type="s", ylim=c(0,1))   #- CDF plot -
```

G.5 pmf and CDF

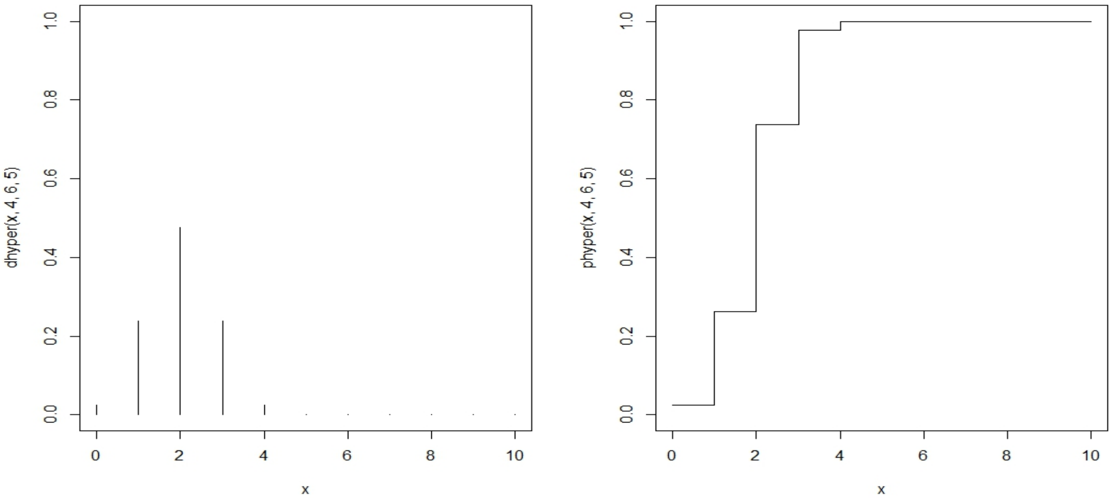

Figure 1: 10 balls, 4 are red. Pick 5 at once. X=number of red picked.

G.6 Ex: Rock sampling

A geologist has collected 10 specimens of basaltic rock and 10 specimens of granite. The geologist instructs a laboratory assistant to randomly select 15 of the specimen for analysis.

a What is the pmf of the number of granite specimen selected for analysis.

b What is the probability hat all specimens of one of the two types of rock are selected for analysis.

c What is the probability that the number of granite specimens selected for analysis is within 1 standard deviation of its mean value?

G.7 Ex: Capture-Recapture

- One popular method of estimating population size of wild animal is called Capture-Recapture method.
- First, you capture m subjects, tag and release them.
- Sometime later, you come back and capture n subjects.
- Within this n subjects, we count how many of them has a tag.
- Let X be the number of tagged subjects.

Our logic is that since

$$E(X) = \frac{nm}{N},$$

It must be that

$$X \approx E(X) = \frac{nm}{N}.$$

If we go with this logic, then our estimator for N will be $\hat{N} = nm/X$.

If we use this estmator in the case $(N = 500, n = 100, m = 100)$, what will be the accuracy of this estimator $\hat{N}$?

Let's define our 'accuracy' as estimator $\hat{N}$ being within 10% of the true value N.

G.8 Cap-Recap Table 1

If N is 500, then P($\hat{N}$ will be within true value $\pm 10\%$) $= P(450 < \hat{N} < 550) = P(a < X < b)$

N = 500

n	m	a	b	$P(a < X < b)$ $= P(450 < \hat{N} < 550)$
10	10	0.18	0.22	0
20	20	0.72	0.88	0
50	50	4.5	5.5	0.19
100	100	18.1	22.2	0.42
200	200	72.7	88.8	0.86
300	300	163.6	200	0.999

G.9 Cap-Recap Table 2

If N is 5000, then P($\hat{N}$ will be within true value $\pm 10\%$) $= P(4500 < \hat{N} < 5500) = P(a < X < b)$

N = 5000

n	m	a	b	$P(a < X < b)$ $= P(4500 < \hat{N} < 5500)$
100	100	1.8	2.2	.28
100	200	3.6	4.4	.20
200	200	7.3	8.9	.15
500	500	45.5	55.6	.57
1000	1000	181.8	222.2	.93

n = number of second round capture

m = number of first round capture-tag-release