Hypergeometric Probability Distribution

Animal Health Risk Analysis
Canadian Food Inspection Agency
3851 Fallowfield Road
Ottawa, Ontario, Canada K2H 8P9

Analyse des risques de la santé des animaux
Agence canadienne d'inspection des aliments
3851, chemin Fallowfield
Ottawa (Ontario) Canada K2H 8P9

Canadian Food Inspection Agency (CFIA)

Agence canadienne d'inspection des aliments (ACIA)

Properties of the Hypergeometric Distribution

Suppose you are selecting a sample of elements from a population and you record whether or not each element possesses a certain characteristic. You are recording the typical success or failure data found in binomial experiments.

If the number of elements in the population is small in relation to the sample size (n / N > 0.05), the probability of a success for a given trial is dependent on the outcomes of preceding trials.
The number x of successes follows what is known as a hypergeometric probability distribution.

It is easy to visualize the hypergeometric random variable X by thinking of a bowl containing D red balls and M - D white balls, for a total of M balls in the bowl.

You select n balls from the bowl and record x, the number if red balls that you see. If you now define a "success" to be a red ball, you have an example of the hypergeometric random variable X.

The hypergeometric distribution applies when, from a finite population consisting of two kinds of objects, a sample of fixed size is drawn successively, without replacement, and we are interested in knowing the number of objects of one kind or the other found in the sample. For example, if a committee studying the economic impact of a new investment interviews 100 people from a small community composed of 350 men and 300 women, then, if X represents the number of women in the sample of interviews, X is a hypergeometric random variable.

Properties of the Hypergeometric Distribution

Let M be the number of objects in a finite population and let D be the number of these objects which are of a particular type, say successes. Therefore, the population contains D successes and M - D failures. Then the probability distribution of the hypergeometric random variable X, which refers to the number of objects which are sucesses in a sample of size n drawn from this population without replacement, or the probability of exactly x successes in a random sample of size n is given as:

$$P(X = x) = \frac{\binom{D}{x}\binom{M-D}{n-x}}{\binom{M}{n}} = \frac{C_x^D C_{n-x}^{M-D}}{C_n^M}$$

for values of x that depend on M, D and n for X = 0, 1, 2, ..., *n* with $C_n^M = \frac{M!}{n!(M-n)!}$

The mean and variance of a hypergeometric random variable are very similar to those of a binomial random variable with a correction for the finite population size.

Properties of the Hypergeometric Distribution

Probability mass function:

$$f(x) = \frac{\binom{D}{x}\binom{M-D}{n-x}}{\binom{M}{n}}$$

Cumulative Distribution function:

$$F(x) = \sum_{i=0}^{x} \frac{\binom{D}{i}\binom{M-D}{n-i}}{\binom{M}{n}}$$

Parameter: n, D, M > 0, n, D, M are all integers, n, D ≤ M

Domain: {0, 1, 2, 3, ..., n}

Mean (μ): $n\frac{D}{M}$

Variance (σ^2): $n\left(\frac{D}{M}\right)\left(\frac{M-D}{M}\right)\left(\frac{M-n}{M-1}\right)$

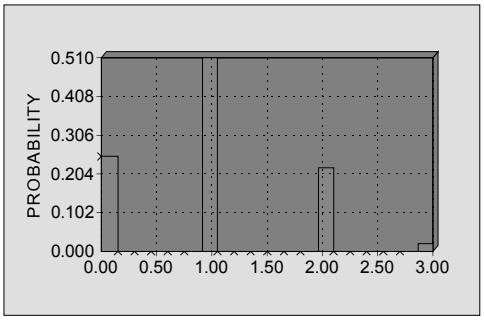
Mode:

Coefficient of skewness (α_3):

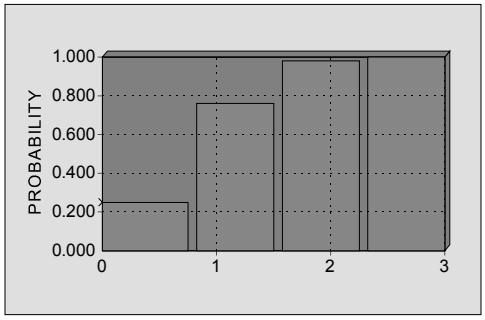
Coefficient of kurtosis (α_4):

Properties of the Hypergeometric Distribution

Probability mass function for H(n = 4, D =3, M = 12) as an output of @RISK simulation



Cumulative distribution function for H(n = 4, D =3, M = 12) as an output of @RISK simullation



Estimating Sample Size Examples

An expression for estimating sample sizes with the hypergeometric approximation was developed by Victor C. Beal Jr. in Regulatory Statistics Part XXII, Hypergeometric Approximation Expanded for Sensitivity.

Sample size (*m*) required to be C% confident of detecting disease in herds/flocks of population size (M) if present at a true prevalence (*p*) and using a test having a sensitivity of Se.
D = the number of animals/birds in the herd/flock that are diseased (*p* x M).

$$m = M - (Se)\left(\frac{D}{2}\right) + 0.5 - (((1 - C)^{\left(\frac{1}{(Se)(D)}\right)}) \times (M - (Se)\left(\frac{D}{2}\right) + 0.5)$$

This expression in Microsoft Excel where C_95 = C (the 95% confidence expressed as a decimal) is presented below. This formula to estimate *m* was used to complete the following six tables employing two values of *p* = 0.01 and 0.05 and three values of Se = 0.80, 0.95, and 1.00.

```
=((M-(Se*(p*M)/2))+0.5-(((1-C_95)^(1/(Se*p*M)))*(M-(Se*(p*M)/2)+0.5)))
```

Estimating Sample Size Examples

Sample size (*m*) required to be 95% confident of detecting disease in herds/flocks of population size (M) if present at a true prevalence (*p* = 0.01) and using a test having a sensitivity of Se = 0.80.

Sensitivity of the test (Se)	0.80	M	m
True prevalence (p)	0.01	100	98
Confidence level (C)	0.95	200	169
		300	213
		400	242
		500	263
		600	278
		700	289
		800	298
		900	305
		1000	311
		1100	316
		1200	321
		1300	324
		1400	327
		1500	330
		1600	333
		1700	335
		1800	337
		1900	339
		2000	340
		3000	351
		4000	356
		5000	359
		6000	362
		7000	363
		8000	364
		9000	365
		10000	366
		20000	370

Estimating Sample Size Examples

Sample size (*m*) required to be 95% confident of detecting disease in herds/flocks of population size (M) if present at a true prevalence (*p* = 0.01) and using a test having a sensitivity of Se = 0.95.

Sensitivity of the test (Se)	0.95	M	m
True prevalence (p)	0.01	100	96
Confidence level (C)	0.95	200	158
		300	195
		400	217
		500	233
		600	244
		700	253
		800	260
		900	265
		1000	269
		1100	273
		1200	276
		1300	279
		1400	281
		1500	283
		1600	285
		1700	287
		1800	288
		1900	289
		2000	290
		3000	298
		4000	302
		5000	304
		6000	306
		7000	307
		8000	308
		9000	308
		10000	309
		20000	311

Estimating Sample Size Examples

Sample size (*m*) required to be 95% confident of detecting disease in herds/flocks of population size (M) if present at a true prevalence (*p* = 0.01) and using a test having a sensitivity of Se = 1.00.

Sensitivity of the test (Se)	1.00	M	m
True prevalence (p)	0.01	100	95
Confidence level (C)	0.95	200	155
		300	189
		400	210
		500	224
		600	235
		700	243
		800	249
		900	254
		1000	258
		1100	261
		1200	264
		1300	266
		1400	268
		1500	270
		1600	272
		1700	273
		1800	275
		1900	276
		2000	277
		3000	284
		4000	287
		5000	289
		6000	291
		7000	292
		8000	293
		9000	293
		10000	294
		20000	296

Estimating Sample Size Examples

Sample size (*m*) required to be 95% confident of detecting disease in herds/flocks of population size (M) if present at a true prevalence (*p* = 0.05) and using a test having a sensitivity of Se = 0.80.

Sensitivity of the test (Se)	0.80	M	m
True prevalence (p)	0.05	100	52
Confidence level (C)	0.95	200	61
		300	65
		400	67
		500	68
		600	69
		700	70
		800	70
		900	70
		1000	71
		1100	71
		1200	71
		1300	71
		1400	71
		1500	72
		1600	72
		1700	72
		1800	72
		1900	72
		2000	72
		3000	72
		4000	73
		5000	73
		6000	73
		7000	73
		8000	73
		9000	73
		10000	73
		20000	73

Estimating Sample Size Examples

Sample size (*m*) required to be 95% confident of detecting disease in herds/flocks of population size (M) if present at a true prevalence (*p* = 0.05) and using a test having a sensitivity of Se = 0.95.

Sensitivity of the test (Se)	0.95	M	m
True prevalence (p)	0.05	100	46
Confidence level (C)	0.95	200	53
		300	56
		400	57
		500	58
		600	58
		700	59
		800	59
		900	59
		1000	60
		1100	60
		1200	60
		1300	60
		1400	60
		1500	60
		1600	60
		1700	60
		1800	61
		1900	61
		2000	61
		3000	61
		4000	61
		5000	61
		6000	61
		7000	61
		8000	61
		9000	61
		10000	61
		20000	61

Estimating Sample Size Examples

Sample size (*m*) required to be 95% confident of detecting disease in herds/flocks of population size (M) if present at a true prevalence (*p* = 0.05) and using a test having a sensitivity of Se = 1.00.

Sensitivity of the test (Se)	1.00	M	m
True prevalence (p)	0.05	100	44
Confidence level (C)	0.95	200	51
		300	53
		400	54
		500	55
		600	56
		700	56
		800	56
		900	57
		1000	57
		1100	57
		1200	57
		1300	57
		1400	57
		1500	57
		1600	57
		1700	57
		1800	57
		1900	58
		2000	58
		3000	58
		4000	58
		5000	58
		6000	58
		7000	58
		8000	58
		9000	58
		10000	58
		20000	58

Ruminant Serum Example

Ruminant serum is imported from bluetongue infected countries in 2 litre bottles. It is then subjected to testing by inoculation of 500 ml of the serum into a susceptible sheep. The sampling of the serum bottles is according to the following scheme and the assumptions that the test has 100% sensitivity and 100% specificity and that the prevalence of infected containors is as high as 5%.

Expression for estimating sample sizes with the hypergeometric approximation
by Victor C. Beal Jr. in Regulatory Statistics Part XXII, Hypergeometric Approximation
Expanded for Sensitivity
n>= M-s*(D/2)+0.5-{((1-C)^(1/(s*(D/2)))*(M-(s*(D/2))+0.5}
Sample size (n) required to be C % confident of detecting infection in shipments of containors of population size (M) if present at a true prevalence (p) and using a test having a sensitivity of s.
D= the number of containors that are infected (p x M)

Sensitivity of the test (s)	**1.00**
True prevalence (p)	**0.05**
Confidence level (C)	**0.95**
Specificity of the test (t)	**1.00**

Ruminant Serum Example

Using the expression above the number of containors to sample (m) in an importation of M containors, is indicated on the right for importations from 20 to 1000 containors.

If 500 ml for sheep inoculation is collected from every 500 litres of imported serum and if the importation comprises 250 containors of 2 litres, what is the probability of detecting an infected importation if only one of the containors is infected.

The sampling scheme indicates that *n* = 52, that is, 52 containors would be sampled (about 10 ml collected from each containor) from M = 250 containors. Here *n*/M = 52/250 ≥ 0.05 and the probability of success in detecting an infected containor in a given trial is dependent on the outcomes of the preceding trials. The number *x* of successes follows a hypergeometric probability distribution.

M	n
20	19
50	34
100	44
150	48
200	51
250	52
300	53
350	54
400	54
450	55
500	55
550	55
600	56
650	56
700	56
750	56
800	56
850	56
900	57
950	57
1000	57

Ruminant Serum Example

Using the hypergeometric distribution with n = 52, M = 250, D = 2 and x = 0, one can find the probability of detecting at least one infected containor as:

$$P(X \geq 1) = 1 - \frac{\binom{D}{x}\binom{M-D}{n-x}}{\binom{M}{n}} = 1 - \frac{\binom{2}{0}\binom{250-2}{52-0}}{\binom{250}{52}}$$

This expression can by computed by spreadsheet software such as Microsoft Excel in which the function is set up as below:

```
=1-HYPGEOMDIST(0,52,2,250)
```

The probability of detecting at least one infected containor in sampling 52 of 250 containors if only 2 containors are infected is therefore P(X≥1) = 0.3734.

Mongooses Example

On a Caribbean island where rabies is endemic in the mongoose population, an effort was made to obtain an estimate of the size of this population. A total of 213 mongooses were trapped and tagged and then released. Then after 6 months, in order to allow sufficient time for mixing of the trapped mongooses with the rest of the population, a second sample of 104 mongooses was taken. Only 13 of the 104 had been previously tagged. Based on these findings, what is the size of the mongoose population.

<u>Solution</u>

For this situation M represents the unknown mongoose population. Letting M' be the estimate of the population, the hypergeometric distribution can be used to estimate the probability of recapturing only 13 out of 213 tagged mongooses, from a sample of 104. With M' set at M' = 500, then D = 213, n = 104 and s = 13.

$$P(X = x) = \frac{\binom{D}{x}\binom{M'-D}{n-x}}{\binom{M'}{n}} = \frac{\binom{213}{13}\binom{500-213}{104-13}}{\binom{500}{104}} = 1.6 \times 10^{-13}$$

Since this is a very small probability, it does not support the estimate of M being equal to 500.

Mongooses Example

This would suggest that as an estimation procedure it would not be unreasonable to select as M' the value that maximizes the likelihood of what has occurred. That is , given D, n and x, maximize:

$$\frac{\binom{D}{x}\binom{M-D}{n-x}}{\binom{M}{n}}$$

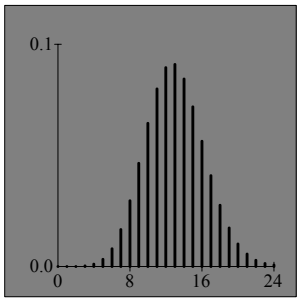
As a function of M. To do this, consider the ratio of the probabilities for two successive values of M:

$$\frac{\binom{D}{x}\binom{M-D}{n-x}}{\binom{M}{n}} \Bigg/ \frac{\binom{D}{x}\binom{M-1-D}{n-x}}{\binom{M-1}{n}} = \frac{(M-D)(M-n)}{M(M-D-n-x)}$$

This ratio is larger than 1 since the probabilities are increasing with M, if and only if (M-D)(M-n) > M(M-D-n-x) or equivalently, if and only if (nD)/x > M. Hence, the population ratio M'/D, is equal to the sample ratio, n/x. For the question asked, M' = (104)(213)/13 =1704.

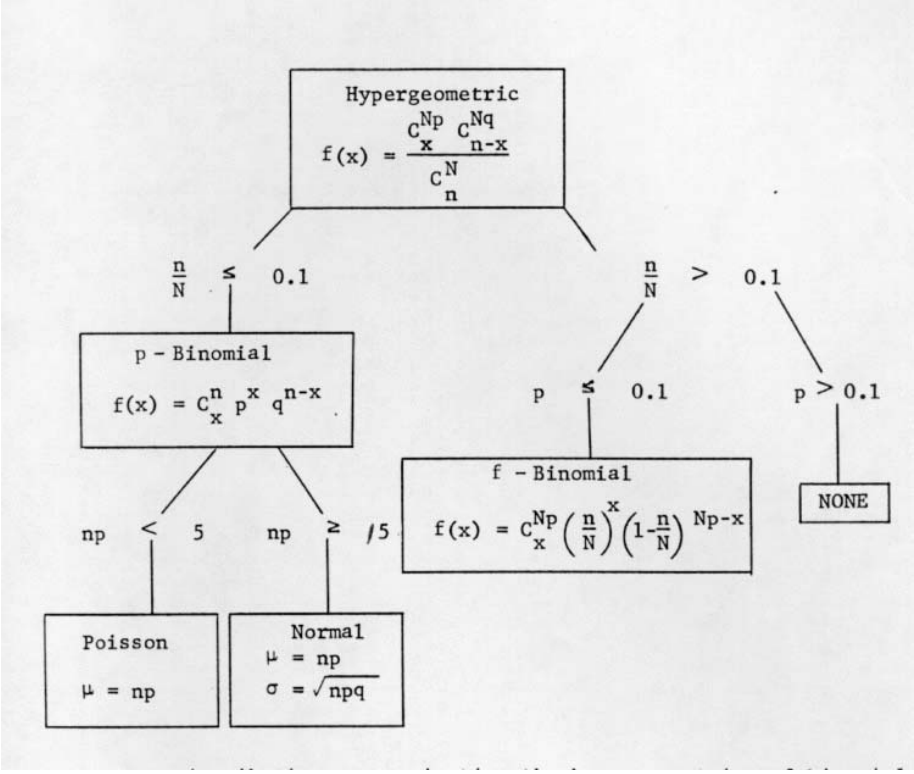
Mongooses Example

The hypergeometic distribution with parameters M=1704, D =213 and n = 104 as portrayed by RiskView (Palisade Corporation) has a mean (μ) and mode both equal to 13 and a variance of 10.687.



The probabilities of the hypergeometic distribution for values where X = x are given at right, calculated by the spreadsheet Microsoft Excel . P(X = 13) =HYPGEOMDIST(13,104,213,1704)

x	P(X=x)
0	5.82E-07
1	9.29E-06
2	7.30E-05
3	3.77E-04
4	1.44E-03
5	4.31E-03
6	0.0106
7	0.0221
8	0.0396
9	0.0620
10	0.0860
11	0.1067
12	0.1194
13	0.1213
14	0.1125
15	0.0958
16	0.0752
17	0.0546
18	0.0368
19	0.0231


Hypergeometric
$f(x) = \frac{C_x^{Np}\, C_{n-x}^{Nq}}{C_n^N}$
$\frac{n}{N} \leq 0.1$
$\frac{n}{N} > 0.1$
p - Binomial
$f(x) = C_x^n\, p^x\, q^{n-x}$
$p \leq 0.1$
$p > 0.1$
f - Binomial
$f(x) = C_x^{Np} \left(\frac{n}{N}\right)^x \left(1-\frac{n}{N}\right)^{Np-x}$
NONE
$np < 5$
$np \geq 5$
Poisson
$\mu = np$
Normal
$\mu = np$
$\sigma = \sqrt{npq}$