HYPERGEOMETRIC DISTRIBUTION

If X is the number of "successes" in a random sample of size n drawn from a population of size N consisting of M "successes", then X has a hypergeometric distribution and

$$P(X=x) = h(x;\, n, M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}.$$

Last time: $0 \le x$ $\quad x \le M$

$n-x \le N-M$ $\quad x \le n$

$n-N+M \le x$

$\max(0,\, n-N+M) \le x$ $\qquad x \le \min(M, n)$

Also: **Mean**: $E(X) = n \cdot \frac{M}{N}$ $\leftarrow$ like the mean np of a binomial rv

Variance: $Var(X) = \underbrace{n \cdot \frac{M}{N}\left(1-\frac{M}{N}\right)}_{\text{like } n\cdot p(1-p)}\left(\frac{N-n}{N-1}\right)$ $\leftarrow$ correction term for sampling without replacement

1. Suppose that in a batch of 20 items, 2 are defective. If 5 of the items are sampled at random:

(a) What is the probability that none of the sampled items are defective?

Let X be hypergeometric rv with $n=5$, $M=2$, $N=20$

$$P(X=0) = \frac{\binom{2}{0}\binom{18}{5}}{\binom{20}{5}} \approx 0.553$$

R: dhyper(0, 2, 18, 5)

(b) What is the probability that exactly 1 of the sampled items are defective?

$$P(X=1) = \frac{\binom{2}{1}\binom{18}{4}}{\binom{20}{5}} \approx 0.395$$

R: dhyper(1,2,18,5)

(c) What is the probability that exactly 3 of the sampled items are defective?

$$P(X=3) = 0 = \frac{\binom{2}{3}\binom{18}{2}}{\binom{20}{5}}$$

zero $\rightarrow \binom{2}{3}$

(d) On average how many defective items will be found in a random sample of 5 items?

$$E(X) = n \cdot \frac{M}{N} = 5 \cdot \frac{2}{20} = 0.5$$

(e) What is the probability that the number of defective items sampled is within 2 standard deviations of the mean number?

$$Var(X) = n\frac{M}{N}\left(1-\frac{M}{N}\right)\left(\frac{N-n}{N-1}\right) = 5\cdot\frac{2}{20}\left(1-\frac{2}{20}\right)\left(\frac{15}{19}\right) \approx 0.355$$

$$\sigma_x = \sqrt{0.355} \approx 0.596$$

R: phyper(1,2,18,5)

$$P(|X-\mu| < 2\sigma_x) = P(-0.7 \le X \le 1.7) = P(X=0 \text{ or } X=1)$$

$$= 0.553 + 0.395 = 0.948$$

R FUNCTIONS:

dhyper(x, M, N-M, n) ← gives $P(X=x)$

phyper(x, M, N-M, n) ← gives $P(X \le x)$

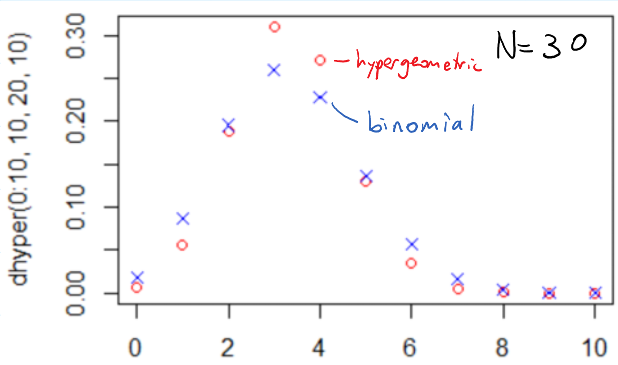
value; successes in pop; failures in pop.; sample size

2. Let X be a hypergeometric random variable with parameters n, M, and N. Let Y be a binomial random variable with parameters n and $p = \frac{M}{N}$. Are $E(X)$ and $E(Y)$ always the same? How does $\text{Var}(Y)$ compare to $\text{Var}(X)$?

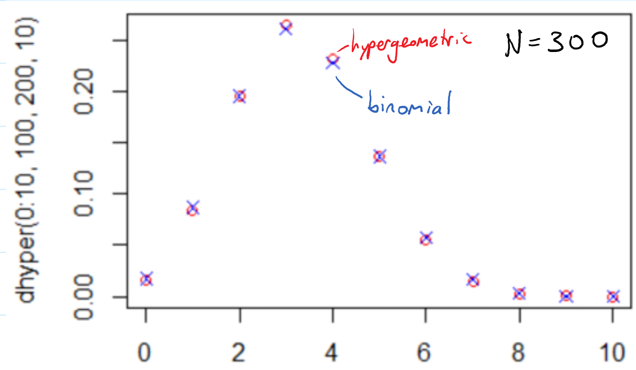
$$E(X) = n\frac{M}{N} = np = E(Y)$$

$$Var(X) = n\cdot\frac{M}{N}\left(1-\frac{M}{N}\right)\left(\frac{N-n}{N-1}\right) = np(1-p)\underbrace{\left(\frac{N-n}{N-1}\right)}_{\le 1} \le np(1-p) = Var(Y)$$

HYPERGEOMETRIC AND BINOMIAL DISTRIBUTIONS




```
plot(0:10, dhyper(0:10, 10, 20, 10), col="red")
points(0:10, dbinom(0:10, 10, 1/3), pch=4, col="blue")
```




```
plot(0:10, dhyper(0:10, 100, 200, 10), col="red")
points(0:10, dbinom(0:10, 10, 1/3), pch=4, col="blue")
```

Why is it called "hypergeometric"?

The ratio of successive probabilities forms a hypergeometric sequence.

ratio of successive terms is a rational function of the index variable

$$\frac{h(x+1;\, n, M, N)}{h(x;\, n, M, N)} = \frac{(M-x)(n-x)}{(x+1)(N-M-n+x+1)}$$

a rational function of x

3. Suppose you have a coin that, when flipped, shows heads with probability p.

(a) You flip the coin repeatedly until the first time the coin shows heads, and then you stop. Let X be the number of coin flips, until (and including) the first flip that shows heads. What is $P(X = 3)$? What is $P(X = k)$?

$$P(X = 3) = (1-p)^2 p$$

T T H

$$P(X = k) = (1-p)^{k-1} p$$

$\underbrace{T\,T \cdots T}_{k-1}\,H$

(b) Let r be a positive integer. Now you flip the coin repeatedly until the coin shows heads r times, and then you stop. Let Y be the number of coin flips, until (and including) the rth flip that shows heads. What is $P(Y = k)$?

$$P(Y = k) = \binom{k-1}{r-1} p^{r-1} (1-p)^{k-r}\, p$$

(r-1) heads in (k-1) flips

_ _ _ _ _ _ _ _ H

k^{th} flip