1. Suppose that in a batch of 20 items, 3 are defective. If 5 of the items are sampled at random:

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

$X \sim \text{Hypergeometric}(5, 3, 20)$

$$P(X=0) = \frac{\binom{3}{0}\binom{17}{5}}{\binom{20}{5}} = \frac{91}{228} \approx 0.399$$

R: dhyper(0, 3, 17, 5)

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

$$P(X=1) = \frac{\binom{3}{1}\binom{17}{4}}{\binom{20}{5}} = \frac{35}{76} \approx 0.461$$

R: dhyper(1, 3, 17, 5)

(c) What is the probability that exactly 4 of the sampled items are defective?

Since the population contains only 3 defective items:

$$P(X=4) = 0 = \frac{\binom{3}{4}\binom{17}{1}}{\binom{20}{5}}$$

This is defined to be zero.

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

Find the mean of X:

$$E(X) = n \cdot \frac{M}{N} = 5 \cdot \frac{3}{20} = \frac{3}{4}$$

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

Find the standard deviation of X:

$$\text{Var}(X) = n \cdot \frac{M}{N}\left(1 - \frac{M}{N}\right)\left(\frac{N-n}{N-1}\right) = 5 \cdot \frac{3}{20}\left(1 - \frac{3}{20}\right)\left(\frac{15}{19}\right) = \frac{153}{304} \approx 0.503$$

$$\sigma_X = \sqrt{\frac{153}{304}} \approx 0.709$$

Now find the requested probability:

$$P(|X - \mu| < 2\sigma_X) = P(-0.668 < X < 2.168)$$

$$= P(X \leq 2)$$

since X takes only nonnegative integer values

$$= \frac{113}{114} \approx 0.991$$

R: phyper(2, 3, 17, 5)

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 $p = \frac{M}{N}$. How does $E(X)$ compare to $E(Y)$? How does $\text{Var}(X)$ compare to $\text{Var}(Y)$?

The means of X and Y are equal:

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

The variance of X is less than the variance of Y if $n > 1$:

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

So $\text{Var}(X) < \text{Var}(Y)$ if $n > 1$.

(If $n = 1$, both X and Y have the same Bernoulli distribution.)

3. Urn 1 contains 100 balls, 10 of which are red. Let X_1 be the number of red balls in a random sample of size 50 from Urn 1. Urn 2 contains 100 balls, 50 of which are red. Let X_2 be the number of red balls in a random sample of size 10 from Urn 2.

(a) Use technology to compute the pmf of X_1. Display the values as a table. Then do the same for the pmf of X_2. What do you notice?

Using Mathematica:

```
In[9]:= Table[PDF[HypergeometricDistribution[50, 10, 100], x], {x, 0, 10}] // N
Out[9]= {0.00059342, 0.00723683, 0.0379933, 0.113096, 0.211413,
  0.259334, 0.211413, 0.113096, 0.0379933, 0.00723683, 0.00059342}
```

] pmf of X_1

```
In[11]:= Table[PDF[HypergeometricDistribution[10, 50, 100], x], {x, 0, 10}] // N
Out[11]= {0.00059342, 0.00723683, 0.0379933, 0.113096, 0.211413,
  0.259334, 0.211413, 0.113096, 0.0379933, 0.00723683, 0.00059342}
```

] pmf of X_2

The two pmfs are the same!

(b) Change the numbers 100, 10, and 50 in this problem and recompute the pmfs of X_1 and X_2. What do you notice?

For example, if $X_1 \sim \text{Hypergeometric}(45, 12, 80)$

and $X_2 \sim \text{Hypergeometric}(12, 45, 80)$

then X_1 and X_2 again have the same pmf.

(c) Make a conjecture about when two hypergeometric random variables have the same pmf.

If $X_1 \sim \text{Hypergeometric}(a, b, N)$ and $X_2 \sim \text{Hypergeometric}(b, a, N)$, then X_1 and X_2 have the same pmf.

4. Suppose that 45% of the phone calls you receive are scam calls. Assume that the probability of a scam call is independent from one call to the next. Let X be the number of calls you receive until (and including) the next scam call.

(a) What is $P(X = 3)$?

$$P(X=3) = \underbrace{(0.55)^2}_{\text{not scam calls}} \underbrace{(0.45)}_{\text{scam call}} \approx 0.136$$

(b) If n is any positive integer, what is $P(X = n)$?

$$P(X=n) = \underbrace{(0.55)^{n-1}}_{\text{not scam calls}} \underbrace{(0.45)}_{\text{scam call}}$$

(c) What is $E(X)$?

$$E(X) = \frac{1}{0.45} \approx 2.22$$

5. Let Y be the number of calls until (and including) the fourth scam call.

(a) What is $P(Y = n)$?

$$P(Y=n) = \underbrace{\binom{n-1}{3}}_{\substack{\text{3 scam calls}\\ \text{in } n-1 \text{ calls}}} \underbrace{(0.45)^3}_{\substack{\text{prob. 3}\\ \text{scam calls}}} \underbrace{(0.55)^{n-4}}_{\substack{\text{prob. } n-4\\ \text{non-scam}\\ \text{calls}}} \underbrace{(0.45)}_{\text{last scam call}} = \binom{n-1}{3}(0.45)^4(0.55)^{n-4}$$

(b) What is $E(Y)$?

$$E(Y) = \frac{4}{0.45} \approx 8.89$$

6. An interviewer must find 10 people who agree to be interviewed. Suppose that when asked, a person agrees to be interviewed with probability 0.4.

(a) What is the probability that the interviewer will have to ask exactly 20 people?

Let $X \sim NB(r=10, p=0.4)$, so $P(X=20) = \binom{20-1}{10-1}(0.4)^{10}(0.6)^{10} \approx 0.0586$

(b) What are the mean and standard deviation of the number of people that the interviewer will have to ask?

$$E(X) = \frac{r}{p} = \frac{10}{0.4} = 25 \qquad Var(X) = \frac{r(1-p)}{p^2} = \frac{10(0.6)}{0.4} = 37.5$$

$$\sigma_X = \sqrt{37.5} \approx 6.12$$

7. If X has a geometric distribution with parameter p, and k is a positive integer, what is $P(X > k)$?

$X > k$ means that the first k trials are all failures.

Thus, $\boxed{P(X>k) = (1-p)^k}$ ← GEOMETRIC TAIL PROBABILITY

8. Scam calls, again. Suppose that 45% of the phone calls you receive are scam calls.

(a) What is the probability that *none* of the first 4 calls are scam calls?

$X \sim \text{Geometric}(0.45) \qquad P(X>4) = (0.55)^4 \approx 0.092$

(b) If none of the first 4 calls are scam calls, what is the probability that none of the first 7 calls are scam calls?

$$P(X>7 \mid X>4) = \frac{P(X>7 \text{ and } X>4)}{P(X>4)} = \frac{P(X>7)}{P(X>4)} = \frac{(0.55)^7}{(0.55)^4} = \underbrace{(0.55)^3}_{\text{This is } P(X>3)} \approx 0.166$$

(c) If none of the first 4 calls are scam calls, what is the probability that none of the first $4 + k$ calls are scam calls?

$$P(X>4+k \mid X>4) = \frac{P(X>4+k)}{P(X>4)} = \frac{(0.55)^{4+k}}{(0.55)^4} = (0.55)^k = P(X>k)$$

OBSERVATION: For $X \sim Geo(p)$ and integers $0<s<t$,

$$\boxed{P(X>t \mid X>s) = P(X>t-s)}$$

← MEMORYLESS PROPERTY of a geometric rv

INTERPRETATION: The waiting time until the next success does not depend on how many failures you have already seen.

BONUS: Show that $\sum_{k=1}^{\infty} k(1-p)^{k-1}p = \frac{1}{p}$. This proves that the mean of a geometric random variable is $\frac{1}{p}$.

Recall the geometric series: $a + ar + ar^2 + ar^3 + \cdots = \frac{a}{1-r}$ for $|r| < 1$

Let $a = r = 1-p$: $\sum_{k=1}^{\infty} (1-p)^k = (1-p) + (1-p)^2 + (1-p)^3 + \cdots = \frac{1-p}{p}$

so: $\sum_{k=1}^{\infty} (1-p)^k = \frac{1}{p} - 1$

Differentiate with respect to p:

$$\sum_{k=1}^{\infty} -k(1-p)^{k-1} = -\frac{1}{p^2}$$

Multiply by $-p$: $\sum_{k=1}^{\infty} k(1-p)^{k-1} p = \frac{1}{p}$