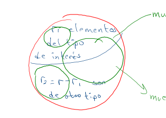
Martes 21 de septiembre de 2021.

2.6 Distribución Hipergeométrica y Distribución de Poisson.

Supón ahora que tenemos una población con r objetos, de los cuales r_1 son del tipo A y r_2 del tipo B. Entonces el número de objetos en la población es $r = r_1 + r_2$.
Se toma una muestra aleatoria sin reemplazo de tamaño n, menor o igual a r, siendo de interés la v.a X: número de objetos del tipo A en la muestra.

Población de tamaño r



muestra aleatoria (m.a) sin reemplazo

muestra aleatoria de tamaño n

rojas ←
verdes
amarillas
negras } otro tipo

X_i : Elemento i-ésimo de la m.a

$$X_i = \begin{cases} 0 & \text{si es de otro tipo} \\ 1 & \text{si es del tipo de interés} \end{cases} \quad \text{Ensayo Bernoulli}$$

Esta v.a se conoce como **hipergeométrica.** Observa que el valor de la v.a no puede ser mayor que n ni mayor que r_1 es decir $x \le \min(r_1, n)$.

X: # de elementos del tipo de interés en la m.a sin reemplazo.

(Se parece a la binomial) No es binomial porque la P(éxito) no es constante porque el muestreo es sin reemplazo (el tamaño del espacio muestral cambia con cada extracción)

$$X = \sum_{i=1}^{n} X_i$$

Ejemplos:

i) $n = 10$ muestra de tamaño 10.
$r_1 = 5$ en la población hay 5 objetos del tipo A.
El número de objetos del tipo A en la muestra a lo más es 5 que es precisamente r_1.

ii) $n = 6$ muestra de tamaño 6.
$r_1 = 7$ en la población hay 7 objetos del tipo A.
El número de objetos del tipo A en la muestra a lo más es 6 que es precisamente el tamaño de la muestra n.

De igual manera el número de objetos del tipo B en la muestra, $n - x$, no puede exceder al número total de objetos del tipo B en la población, es decir, $n - x \le r_2 \;\therefore\; n - r_2 \le x$ y tampoco puede ser un número negativo por lo tanto $\max(0, n - r_2) \le x \le \min(r_1, n)$.

	x
Bernoulli	0, 1
Binomial	0, 1, 2, ..., n
Geométrica	1, 2, 3, ...
Hipergeométrica	max (0, n - r2), ..., min (n, r1)

2 blancas, 5 otro color → sin reemplazo

$P(b_1) = \frac{2}{7}$

$P(b_2) = P(b_2 \mid b_1) + P(b_2 \mid b_1')$

$= \frac{1}{6} + \frac{2}{6} = \frac{3}{6} \neq \frac{2}{7}$

Ejemplos:

i) $n = 10$ muestra de tamaño 10.
$r_1 = 5$ en la población hay 5 objetos del tipo A.
El número de objetos del tipo A en la muestra a lo más es 5 que es precisamente r_1.

ii) $n = 6$ muestra de tamaño 6.
$r_1 = 7$ en la población hay 7 objetos del tipo A.
El número de objetos del tipo A en la muestra a lo más es 6 que es precisamente el tamaño de la muestra n.

2 blancas, 5 otro → Con reemplazo

$P(b_1) = \frac{2}{7}$

$P(b_2) = \frac{2}{7}$ etc.

X : # de elementos del tipo de interés en la m.a

Obs.- Este experimento se puede hacer

i) Con reemplazo → la v.a es binomial

ii) Sin reemplazo → la v.a es hipergeométrica.

De igual manera el número de objetos del tipo B en la muestra, $n - x$, no puede exceder al número total de objetos del tipo B en la población, es decir, $n - x \le r_2 \;\therefore\; n - r_2 \le x$ y tampoco puede ser un número negativo por lo tanto $\max(0, n - r_2) \le x \le \min(r_1, n)$.

Ejemplos:

i) $n = 6$ muestra de tamaño 6.

$r_1 = 7$ en la población hay 7 objetos del tipo A.
$r_2 = 5$ y 5 objetos del tipo B, entonces el número total de elementos en la población es $r = 12$.
La variable aleatoria de interés toma valores entre 1 y 6 inclusive, ya que $\max(0, 6-5) \le x \le \min(7, 6)$.

$x = 1, 2, ..., 6$

$\max(0, 1) \le x \le \min(7, 6)$
$1 \le x \le \min(7, 6)$

ii) $n = 6$ muestra de tamaño 6.
$r_1 = 5$ en la población hay 5 objetos del tipo A.
$r_2 = 10$ elementos del tipo B, entonces el número total de elementos en la población es $r = 15$.
La variable aleatoria de interés toma valores entre 0 y 5 inclusive, ya que $\max(0, 6-10) \le x \le \min(5, 6)$.

0, -4

$x = 0, 1, ..., 5$

Si tomamos una muestra aleatoria de tamaño 6, el mínimo número de objetos de la clase A es cero (hay 10 objetos de la clase B, entonces los 6 de la muestra pueden ser de este tipo) y el máximo número posible es 5 ya que sólo hay 5 del tipo A en la población.

Función de probabilidad.
La expresión $X \sim H(x; r_1, r_2, n)$ indica que la variable aleatoria X tiene una distribución hipergeométrica con parámetros r_1, r_2 y n, tal distribución es de la forma:

$$f_X(x) = \frac{\binom{r_1}{x}\binom{r - r_1}{n - x}}{\binom{r}{n}} \qquad \max(0, n - r_2) \le x \le \min(r_1, n)$$

$\frac{\binom{r_1}{x}\binom{r_2}{n-x}}{\binom{r}{n}}$ $\qquad r = r_1 + r_2$

Recordemos que NO HAY INDEPENDENCIA.

Esperanza y varianza.

Sean las variables aleatorias Bernoulli $X_i = \begin{cases} 1 & \text{si } X_i \text{ es del tipo A} \\ 0 & \text{si } X_i \text{ es del tipo B} \end{cases}$ para $i = 1, 2, \ldots, n$.

Para las cuales $E(X_i) = p = \frac{r_1}{r}$.

Observa que la v.a. hipergeométrica es la suma de estas variables aleatorias Bernoulli, $X = \sum_{i=1}^{n} X_i$, usando las propiedades del valor esperado se tiene que:

$$E(X) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = nE(X_i)$$

Y así se tiene que la esperanza de la v.a hipergeométrica es: $\boxed{E(X) = n\frac{r_1}{r}}$

La muestra se toma sin reemplazo, por lo tanto las variables aleatorias X_i no son independientes. Para encontrar la varianza de la v.a hipergeométrica necesitamos la varianza de cada X_i y la covarianza $Cov(X_i, X_j)\ \forall\ i < j$.

La varianza de cada Bernoulli es $V(X_i) = pq = \left(\frac{r_1}{r}\right)\left(1 - \frac{r_1}{r}\right)$ (1)

Y la covarianza entre X_i y X_j está dada por:
$Cov(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)\ \forall\ i < j$

Recuerda que: $E(g(x)) = \sum_x g(x) P(X = x)$

$E(X_i X_j) = \sum_{x_i x_j} x_i x_j P(X_i X_j = x_i x_j)$

Los valores posibles que toma el producto son 0 y 1 de la manera siguiente:

$$\left.\begin{aligned} X_i = 0 \ &\text{y}\ X_j = 1 \\ X_i = 1 \ &\text{y}\ X_j = 0 \\ X_i = 0 \ &\text{y}\ X_j = 0 \end{aligned}\right\} X_i X_j = 0$$

$X_i = 1$ y $X_j = 1 \quad X_i X_j = 1$

Por lo tanto $= 1\,P(X_i X_j = 1)$ Ley de la multiplicación

$$E(X_i X_j) = 1P(X_i = 1, X_j = 1) = P(X_i = 1)P(X_j = 1 \mid X_i = 1) = \frac{r_1}{r}\frac{r_1 - 1}{r - 1}$$

Entonces $Cov(X_i, X_j) = \frac{r_1}{r}\frac{r_1 - 1}{r - 1} - \left(\frac{r_1}{r}\right)^2 = \frac{r_1}{r}\left(\frac{r_1 - 1}{r - 1} - \frac{r_1}{r}\right) = \frac{r_1}{r}\left(\frac{\cancel{r_1 r} - r - \cancel{r_1 r} + r_1}{r(r-1)}\right) = \frac{r_1}{r}\left(\frac{r_1 - r}{r(r-1)}\right)$ (2)

Usando las ecuaciones (1) y (2) en la expresión de la varianza:

$$V(X) = V\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} V(X_i) + 2\sum_{i<j} Cov(X_i, X_j)$$

$$= n\frac{r_1}{r}\left(1 - \frac{r_1}{r}\right) + \cancel{2}\frac{n(n-1)}{\cancel{2}}\frac{r_1}{r}\left(\frac{r_1 - r}{r(r-1)}\right)$$

$$= n\frac{r_1}{r}\left(1 - \frac{r_1}{r}\right) + n(n-1)\frac{r_1}{r}\left(\frac{r_1 - r}{r(r-1)}\right)$$

$$= n\frac{r_1}{r}\left(\frac{r - r_1}{r}\right) - n\left(\frac{r_1}{r}\right)\left(\frac{r - r_1}{r}\right)\left(\frac{n-1}{r-1}\right)$$

$$= n\frac{r_1}{r}\left(\frac{r - r_1}{r}\right)\left(1 - \frac{n-1}{r-1}\right)$$

$\binom{n}{2} = \frac{n!}{2(n-2)!} = \frac{n(n-1)\cancel{(n-2)!}}{2\cancel{(n-2)!}}$

Finalmente tenemos que la varianza de la v.a hipergeometrica está dada por:

$$\boxed{V(X) = n\left(\frac{r_1}{r}\right)\left(1 - \frac{r_1}{r}\right)\left(1 - \frac{n-1}{r-1}\right)}$$

Nota._ Supón que se tiene una población de tamaño r con r_1 objetos de la clase A y $r - r_1$ de la clase B, y se toma una muestra aleatoria (m.a) de tamaño n. La v.a Y: Número de objetos de la clase A en la muestra puede tener dos distribuciones de probabilidad diferentes, dependiendo de la manera en que se toma la m.a:

i) Si la muestra se toma con reemplazo, la probabilidad de éxito (el objeto es de la clase A) es constante igual a $\frac{r_1}{r}$, hay independencia y por lo tanto $Y \sim b(y; n, p)$ con $E(X) = np = n\frac{r_1}{r}$ y

$$V(Y) = npq = n\frac{r_1}{r}\left(1 - \frac{r_1}{r}\right)$$

ii) Si la muestra se toma sin reemplazo entonces $Y \sim H(y; r_1, r_2, n)$ con $E(Y) = np = n\frac{r_1}{r}$ y

$$V(Y) = n\left(\frac{r_1}{r}\right)\left(1 - \frac{r_1}{r}\right)\left(1 - \frac{n-1}{r-1}\right)$$

El valor esperado es el mismo pero si el muestreo es sin reemplazo la varianza es menor ya que:

$$n\left(\frac{r_1}{r}\right)\left(1 - \frac{r_1}{r}\right) > n\left(\frac{r_1}{r}\right)\left(1 - \frac{r_1}{r}\right)\left(1 - \frac{n-1}{r-1}\right)$$

pero $1 > 1 - \frac{n-1}{r-1}$.

Nota.- Entonces es mejor un muestreo sin reemplazo porque tiene menor varianza.

Si r es grande comparado con n, entonces $\frac{n}{r} \to 0$ y la varianza de la hipergeométrica se parece mucho a la varianza de la binomial, es decir para n fijo y r grande (poblaciones muy grandes) hay poca diferencia entre muestreo con y sin reemplazo (binomial e hipergeométrica).

Función de probabilidad acumulada.

Existen en escuelas del IPN tablas de la distribución acumulada realizadas por los profesores, pero en los libros difícilmente se encontraran debido a la diversidad de casos, en general esta probabilidad se puede encontrar usando la definición de acumulada y la f.d.p de la Hipergeométrica como sigue:

$$F_X(x) = P(X \le x) = \sum_{i=\max(0, n-r_2)}^{x} \frac{\binom{r_1}{i}\binom{r - r_1}{n - i}}{\binom{r}{n}}$$

Ejemplos:

1._ Supón que una urna contiene cinco bolas rojas y diez azules. Si se seleccionan al azar sin reemplazo siete bolas.
a) ¿Cuál es la probabilidad de obtener al menos tres bolas rojas?
b) Si $\overline{X}$ representa la proporción de bolas rojas en la muestra ¿Cuáles son la media y la varianza de $\overline{X}$.

Solución

$r_1 = 5$ bolas rojas.
$r_2 = 10$ bolas azules.
$n = 7$ tamaño de la muestra.
$r = 15$ total de bolas en la urna.
X: Número de bolas rojas en la muestra.
$X \sim H(x; r_1 = 5, r_2 = 10, n = 7)$

$\max(0, n - r_2) \le x \le \min(n, r_1)$
$\max(0, 7 - 10) \le x \le \min(7, 5)$
$x = 0, 1, 2, ..., 5.$

$P(X \ge 3) = P(X = 3) + P(X = 4) + P(X = 5)$

$$= \frac{\binom{5}{3}\binom{10}{4} + \binom{5}{4}\binom{10}{3} + \binom{5}{5}\binom{10}{2}}{\binom{15}{7}} = \frac{2100 + 600 + 45}{6435} = \frac{2745}{6435} = \frac{61}{143} = 0.4265$$

b) Observa que $\overline{X} = \frac{X}{n}$ por lo tanto $E(\overline{X}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X)$ y $V(\overline{X}) = V\left(\frac{X}{n}\right) = \frac{1}{n^2}V(X)$

$$E(\overline{X}) = \frac{1}{7}\left(\frac{7(5)}{15}\right) = \frac{1}{3}$$

$E(\frac{1}{n}X)$

$$V(\overline{X}) = \left(\frac{1}{7}\right)^2\left(\frac{7(5)}{15}\right)\left(1 - \frac{5}{15}\right)\left(1 - \frac{6}{14}\right) = \frac{10(8)}{7(3)(15)(14)} = \frac{8}{441} = 0.018$$

2._ En un almacén se tienen 10 impresoras de las cuales 4 son defectuosas. Una compañía selecciona 5 de las máquinas al azar.
a) ¿Cuál es la probabilidad de que las 5 máquinas sean no defectuosas?
b) La compañía repara las impresoras defectuosas a un costo de 50 dólares cada una. Encuentra la media y la varianza del costo total de reparación.

Solución

$r_1 = 6$ máquinas buenas.
$r_2 = 4$ máquinas defectuosas.
$n = 5$ tamaño de la muestra.
$r = 10$ total de máquinas.

X: Número de máquinas buenas en la muestra.

$X \sim H(x; r_1 = 6, r_2 = 4, n = 5)$

$C(X) = 50(5 - X) = 250 - 50X$

$E(C(X)) = E(250 - 50X)$

$= 250 - 5\cancel{0}(5)\frac{6}{1\cancel{0}} = 250 - 150 = 100$

$\therefore$ # de máquinas defectuosas en la m.a es $\underbrace{5 - X}$ por cada una de estas me toca pagar \$50.00

a) $P(X = 5) = \frac{\binom{6}{5}\binom{4}{0}}{\binom{10}{5}} = \frac{6}{252} = \frac{1}{42} = 0.024$

b) Sea X : # de máquinas defectuosas en la m.a

b) Ahora sea la v.a de interés Y: Número de máquinas defectuosas en la muestra.

$Y \sim H(y; r_1 = 4, r_2 = 6, n = 5)$
El costo de reparación está dado como $C(Y) = 50Y$.
Por lo tanto:

$$E[C(Y)] = 50E(Y) = 50(5)\frac{4}{10} = 100$$

$$V[C(Y)] = (50)^2 V(Y) = (50)^2(5)\left(\frac{4}{10}\right)\left(\frac{6}{10}\right)\left(1 - \frac{4}{9}\right) = 1{,}666.66$$

3._ Una corporación muestrea sin reemplazo $n = 3$ empresas para adquirir ciertos suministros. La muestra se selecciona de un conjunto de 6 empresas, de las cuales cuatro son locales y dos no lo son. Sea Y el número de empresas foráneas entre las tres escogidas. Encuentra las siguientes probabilidades:
a) $P(Y = 1)$
b) $P(Y \ge 1)$
c) $P(Y \le 1)$

Solución

$r_1 = 2$ empresas foráneas.
$r_2 = 4$ $r_2 = 4$ empresas locales.
$n = 3$ tamaño de la muestra.
$r = 6$ total de empresas.

Y: Número de empresas foráneas.
$Y \sim H(y; r_1 = 2, r_2 = 4, n = 3)$

a) $P(Y = 1) = \frac{\binom{2}{1}\binom{4}{2}}{\binom{6}{3}} = \frac{3}{5} = 0.6$

b) $P(Y \ge 1) = 1 - P(Y = 0) = 1 - \frac{\binom{2}{0}\binom{4}{3}}{\binom{6}{3}} = 1 - \frac{1}{5} = 0.8$

c) $P(Y \le 1) = P(Y = 0) + P(Y = 1) = \frac{1}{5} + \frac{3}{5} = \frac{4}{5} = 0.8$