Inferencia Estadística

Dra. Graciela González Farías
Dr. Ulises Márquez

Maestría en Cómputo Estadístico.

CIMAT Monterrey.

Agradecimientos

En forma de agradecimiento, se enlistan personas que han contribuido de una u otra forma en la construcción de estas notas a través de los años:

- Víctor Muñiz
- Juan Antonio López
- Sigfrido Iglesias González
- Rodrigo Macías Paéz
- Edgar Jiménez
- Todos los estudiantes que han colaborado con sugerencias y comentarios sobre estas notas.

Estas notas son de uso exclusivo para enseñanza y no pretende la sustitución de los textos y artículos involucrados.

Temario

1. Variables aleatorias y distribuciones de probabilidad.
    a) Distribuciones de probabilidad de variables aleatorias discretas.
    b) Procesos de Poisson.
    c) Distribuciones de probabilidad de variables aleatorias continuas.
    d) Métodos gráficos para la identificación de distribuciones.
    e) Estimación de densidades.
    f) Distribuciones de probabilidad de vectores aleatorios.
    g) Esperanzas condicionales y regresión.
    h) Modelos jerárquicos, compuestos y mezclas de variables aleatorias.
    i) Transformaciones de variables aleatorias.
    j) Simulación de variables aleatorias.
    k) Convergencia de variables aleatorias y el Teorema del Límite Central.

Temario

2. Distribuciones muestrales y métodos de estimación.
   a) Propiedades de los estimadores.
   b) Estimadores no insesgados
   c) Distribuciones muestrales.
   d) Principio de máxima verosimilitud.
   e) Estimación puntual.
   f) Bootstrap y jackknife.

Temario

3. Pruebas de Hipótesis e intervalos de confianza.
   a) Definición de conceptos.
   b) Potencia de la prueba.
   c) Pruebas para dos poblaciones normales independientes.
   d) Pruebas para medias en muestras pareadas.
   e) Pruebas básicas de varianzas.
   f) Pruebas para proporciones.
   g) Conceptos de estimación bayesiana.
   h) Temas optativos de modelos para presentaciones finales, por ejemplo:
      1. Pruebas no-paramétricas clásicas.
      2. Pruebas de permutaciones.
      3. Estimación no paramétrica (suavizadores y splines).
      4. Modelos gráficos probabilístas.
      5. Entre muchos otros.

Evaluación y acreditación

- Dos exámenes parciales, 18 de septiembre y 6 de noviembre: 15 %, cada uno.
- Evaluación de las tareas (de 2 tipos) y actividades en clase y asistencia: 40 %.
- Un examen final, consistente en una exposición donde se entrega un reporte y se hace una presentación de 1/2 hora. La presentación debe incluir antecedentes, metodología, un ejemplo práctico y compartir el código. Deberán entregar a los instructores y a sus compañeros el resumen. Adicionalmente, deberán dejar un ejercicio sobre el tema a sus compañeros que calificarán en forma honesta: 30 %.

Las tareas tienen una frecuencia quincenal e incluyen TODOS los ejercicios dejados en las notas y requerirán en general el uso de recursos computacionales.

Textos

- **Larry Wasserman (2004) . All of Statistics, A concise course in Statistical Inference. Springer.**
- F.M. Dekking, C. Kraaikamp, H.P. Lopuhaa L.E. Meester (2005). A Modern Introduction to Probability and Statistics, Understanding Why and How. Springer text in Statistics.
- John A. Rice (1995). Mathematical Statistics and Data Analysis, Second Edition. Duxbury Press.
- Casella & Berger. (2002). Statistical Inference, Second Edition . Duxbury Press.
- Richard J. Larsen and Morris L. Marx (2011). An Introduction to Mathematical Statistics and its Applications. Fifth Edition. Prentice Hall.

Modelos probabilísticos

Modelos probabilísticos

Recordemos que una variable aleatoria discreta es aquella que sólo toma un número contable de valores (finito o infinito). Por ejemplo:

- El número de plantas con daños visibles producidos por una plaga.
- El número de individuos a favor de un partido político.
- El número de televisores con defectos en su selector de canales de un lote de 100 televisores.
- El número de personas en la fila en un centro de servicio al público, entre las 9 y 10 de la mañana.

Notamos que en cada una de esas situaciones uno lleva a cabo algún tipo de **conteo**.

Modelos probabilísticos

En un principio uno debería:

- Examinar cada caso;
- Ver cuáles son las condiciones específicas en que se realiza el muestreo;
- Establecer los supuestos de simplicidad que sean factibles; y,
- Determinar el modelo probabilístico que mejor describa el comportamiento de la característica bajo estudio (verificando su validez).

Modelos probabilísticos

Este procedimiento general ha dado lugar a un cierto número de modelos que aparecen frecuentemente en las aplicaciones. Así, lo que haremos aquí, es construir estos modelos particulares formando un catálogo básico que nos permita referenciar nuestras situaciones particulares a alguno de estos. En la construcción del catálogo, contemplamos varios puntos:

1. Supuestos necesarios para identificar el uso del modelo.
2. Construcción del modelo: función de probabilidad y de probabilidad acumulada.
3. Momentos: Media, Varianza, Función generatriz de momentos (cuando aplique).

Modelos de variables discretas

Distribución uniforme discreta

Cuando se tiene un número **finito** de resultados de un experimento, $x_1, x_2, \ldots, x_n$, cada uno de ellos **igualmente probable**, se dice que se tiene una variable aleatoria que puede ser modelada por una distribución uniforme discreta. Notación:

$$X \sim U(\, x_1, x_2, \ldots, x_n\,).$$

Se lee: X se distribuye como una variable aleatoria uniforme.

Este tipo de modelo aparece típicamente en la selección de **muestras al azar**.

El único parámetro de la distribución es n, el número posible de resultados.

Distribución uniforme discreta

Como la suma de todas las probabilidades debe ser uno y todos los valores son equiprobables, entonces la función de masa está dada por

$$f(x) = \begin{cases} 1/n & \text{para } x = x_1, x_2, \ldots, x_n \\ 0 & \text{en otro caso} \end{cases}$$

Bajo esta definición, claramente $f(x) \geq 0$ para toda x, y

$$\sum_{i=1}^{n} f(x_i) = n\left(\frac{1}{n}\right) = 1.$$

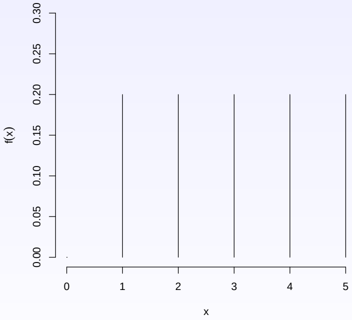
El caso más común de esta distribución es cuando la variable toma los valores enteros

$$\{1, 2, \ldots, n\}.$$

El resto de los resultados los estableceremos para cuando la variable aleatoria uniforme toma estos valores.

Distribución uniforme discreta

El parámetro (valor que identifica unívocamente al modelo) de la distribución es n, el número total de objetos. Se dice en este caso que se trata de un espacio de probabilidad equiprobable.



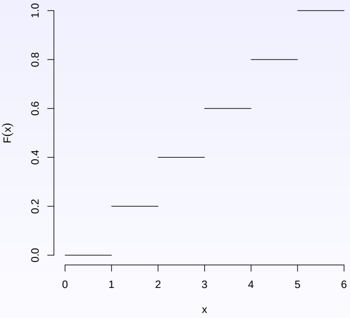


Figura: Distribución uniforme para $n = 5$.

Distribución uniforme discreta

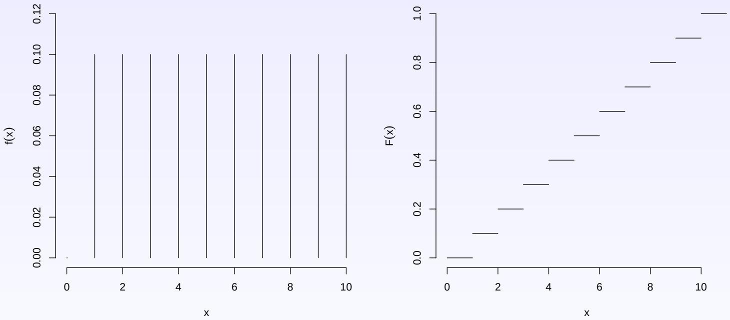


Figura: Distribución uniforme $n = 10$.

Distribución uniforme discreta

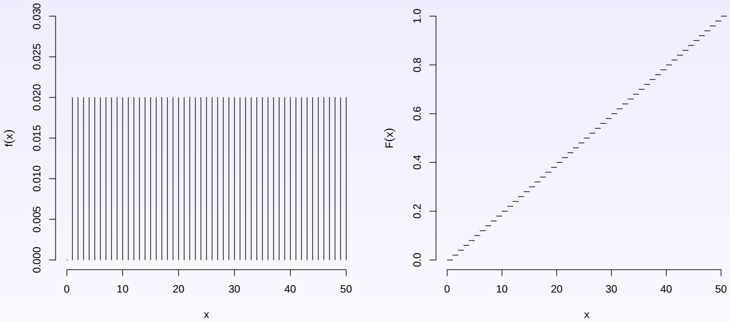


Figura: Distribución uniforme $n = 50$.

Distribución uniforme discreta

Media y Varianza:

$$E(X) = \sum_{x=1}^{n} x\frac{1}{n} = \frac{1+2+\cdots+n}{n} = \frac{n+1}{2},$$

y

$$\begin{aligned}
V(X) = \; & E(X^2) - [E(X)]^2 = \sum_{x=1}^{n} x^2\frac{1}{n} - \left(\frac{n+1}{2}\right)^2 \\
= \; & \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 = \frac{2n^2+3n+1}{6} - \frac{n^2+2n+1}{4} \\
= \; & \frac{2(2n^2+3n+1)-3(n^2+2n+1)}{12} = \frac{4n^2+6n+2-3n^2-6n-3}{12} \\
= \; & \frac{n^2-1}{12}.
\end{aligned}$$

Distribución uniforme discreta

<u>**Función Generatriz de Momentos**</u>:

$$M_X(t) = \mathrm{E}\left(e^{tX}\right) = \sum_{x=1}^{n} e^{tx} \cdot f(x) = \sum_{x=1}^{n} e^{tx} \cdot \frac{1}{n}$$

$$= \frac{1}{n} \sum_{x=1}^{n} \left(e^{t}\right)^{x} = \frac{1}{n}\left[\frac{e^{t} - (e^{t})^{n+1}}{1 - e^{t}}\right] = \frac{1}{n} \cdot \frac{e^{t}(1 - e^{nt})}{1 - e^{t}}, \quad \forall t.$$

Nota: $M_X(0) = 0$, aplicando la regla de L'Hopital.

Se puede consultar más información sobre la función característica en:
`en.wikipedia.org/wiki/Characteristic_function_(probability_the`

Distribución Bernoulli

Definamos un experimento en el cual hay únicamente dos posibles resultados: a uno de ellos le llamamos "éxito", al otro "fracaso". Este tipo de variables aparece frecuentemente en nuestros conteos, por ejemplo, si pensamos en clasificar nuestros productos como: defectuoso, no defectuoso; grande o pequeño; azul o blanco; sí o no, etc.

Generalmente le asignamos un valor de 1 al "éxito" y un valor de 0 al "fracaso". Notemos que la asignación de los valores numéricos es arbitraria. Este procedimiento sirve de base para la construcción de otras distribuciones de gran utilidad.

Distribución Bernoulli

Definamos $X = \#$ de "éxitos", entonces

$$x = \begin{cases} 1 & \text{si } X = \text{"éxito"}, \text{ esto con probabilidad } p \\ 0 & \text{si } X = \text{"fracaso"}, \text{ esto con probabilidad } (1-p) = q \end{cases}$$

o bien,

$$f(x) = p^x(1-p)^{1-x}, \quad \text{para} \quad x = 0, 1; \qquad \text{(Distribución Bernoulli).}$$

Notación: $X \sim Bernoulli(p)$. El único parámetro de la distribución de ésta variable aleatoria es p, la probabilidad de éxito.

Distribución Bernoulli

Por ejemplo:

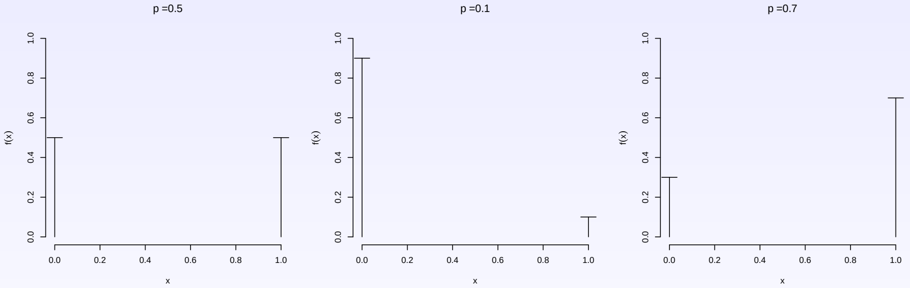


Figura: Función de densidad para $X \sim Bernoulli(p)$.

El experimento que guió al modelo, dos posibles resultados con probabilidades p y q $(p + q = 1)$, se denomina experimento Bernoulli.

Distribución Bernoulli

Media y Varianza:

$$
\begin{aligned}
\mathrm{E}(X) &= \sum_{x=0}^{1} x \cdot f(x) = 0 \cdot (1-p) + 1 \cdot p = p \\
\mathrm{E}(X^2) &= \sum_{x=0}^{1} x^2 \cdot f(x) = 0^2 \cdot (1-p) + 1^2 \cdot p = p \\
\mathrm{V}(X) &= \mathrm{E}(X^2) + [\mathrm{E}(X)]^2 = p - p^2 = p(1-p) = pq
\end{aligned}
$$

Función Generatriz de Momentos:

$$
\begin{aligned}
M_X(t) &= \mathrm{E}\left(e^{tX}\right) = \sum_{x=0}^{1} e^{tx} \cdot f(x) = e^{t \cdot 0}(1-p) + e^{t \cdot 1} p \\
&= (1-p) + pe^t = q + pe^t, \quad \forall t.
\end{aligned}
$$

Distribución Binomial e Hipergeométrica

Para las siguentes distribuciones definiremos nuestra variable aleatoria de la siguiente manera:

$$X = \# \text{ de "éxitos" en } n \text{ observaciones.}$$

X es el número de unos o "éxitos" que se presenten en la muestra de tamaño n. Lo que asigna un patrón distinto de comportamiento, es la **forma como se realizan (condiciones) las n observaciones**.

Distribución Binomial

Supongamos que podemos hacer:

1. n observaciones independientes.
2. la probabilidad de éxito en cada observación permanece constante, esto es, siempre es p.

En otras palabras, estamos asumiendo que nuestra población es suficientemente grande como para tomarla como "infinita" y que podemos garantizar la independencia entre observaciones, esto es, no obtenemos información adicional para predecir el siguiente resultado sólo porque ya observamos al (los) anterior (es). Esto refleja un comportamiento de procesos que lo podríamos denominar estable. Estas condiciones deben estar presentes al menos durante el período en que se realiza el estudio.

Distribución Binomial

Por ejemplo, pensemos en un proceso industrial, producción de mangueras para gas. Las clasificaremos como defectuosas (éxito) o no defectuosas, de acuerdo a si cumplen o no con el tamaño requerido. Se toma una muestra de tamaño tres y se hacen las mediciones de cada una de las mangueras.

- Asumimos que nuestra población son todas las mangueras que pasan por ese proceso (# muy grande),
- Se asume que no hay ninguna causa que motive desperfectos sistemáticos,
- También notemos que, bajo estas condiciones, la probabilidad de que una manguera no sea de las medidas requeridas deberá permanecer constante para las 3 observaciones y, más aún, esta probabilidad está dada por otro mecanismo independiente al conteo que nos atañe en este momento.

Distribución Binomial

Por ejemplo, p se podría determinar como una proporción observada a través del tiempo, o bien porque conozcamos la ley de fallas en cortes, de la maquinaria que se emplea. Para nuestros propósitos, p es un valor dado.

Notemos primero que el número posible de valores que puede tomar X es $\{1, 2, \ldots, n\}$. Ahora obtengamos la ley de comportamiento asociada con esta variable:

$$\begin{aligned} \mathrm{P}(X = 0) &= \mathrm{P}(\text{ninguna manguera defectuosa en la muestra de } n = 3) \\ &= \mathrm{P}(X_1 = 0, X_2 = 0, X_3 = 0) \end{aligned}$$

donde X_i denota la i-ésima observación. Cabe notar que cada X_i toma el valor de 0 ó 1 con probabilidades constantes q y p, respectivamente. Esto es, cada X_i **representa una variable Bernoulli**(p) **y son independientes entre sí**.

Distribución Binomial

Entonces:

$$\begin{aligned} \mathrm{P}(X_1 = 0, X_2 = 0, X_3 = 0) = & \ \mathrm{P}(X_1 = 0)\mathrm{P}(X_2 = 0)\mathrm{P}(X_3 = 0) \\ = & \ qqq = q^3 = \binom{3}{0} q^3 \end{aligned}$$

A manera de ejercicio, puedes verificar que:

$$\begin{aligned} \mathrm{P}(X = 1) = & \ \mathrm{P}(\{X_1 = 1, X_2 = 0, X_3 = 0\} \text{ ó } \{X_1 = 0, X_2 = 1, X_3 = 0\} \text{ ó } \{X_1 = 0, X_2 = 0, X_3 = 1\}) \\ = & \ 3pqq = 3pq^2 = \binom{3}{1} pq^2, \qquad \mathbf{1} \ \textbf{“éxito”}; \\ \mathrm{P}(X = 2) = & \ \mathrm{P}(\{X_1 = 1, X_2 = 1, X_3 = 0\} \text{ ó } \{X_1 = 1, X_2 = 0, X_3 = 1\} \text{ ó } \{X_1 = 0, X_2 = 1, X_3 = 1\}) \\ = & \ 3ppq = 3p^2q = \binom{3}{2} p^2q, \qquad \mathbf{2} \ \textbf{“éxitos”}; \\ \mathrm{P}(X = 3) = & \ \mathrm{P}(\{X_1 = 1, X_2 = 1, X_3 = 1\}) \\ = & \ ppp = 3p^3 = \binom{3}{3} p^3, \qquad \mathbf{3} \ \textbf{“éxitos”}. \end{aligned}$$

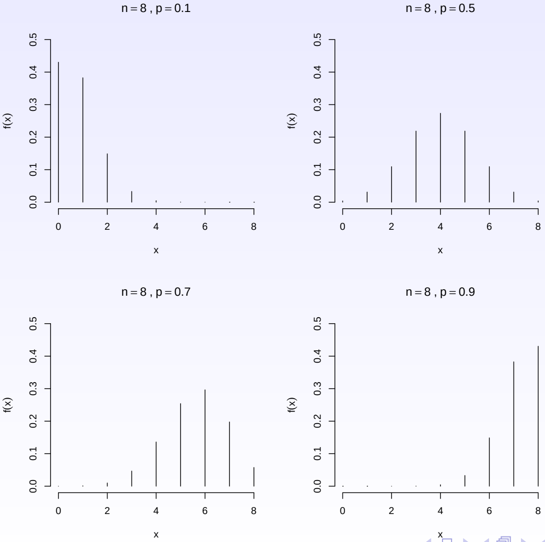
Distribución Binomial

Esto es, tenemos la Distribución Binomial

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{para } x = 0, 1, \ldots, n \\ \\ 0 & \text{en otro caso} \end{cases}$$

En este caso se dice que X sigue una distribución binomial con parámetros (n, p). Notación: $X \sim B(n, p)$.

Distribución Binomial



Distribución Binomial

Se puede verificar que la suma, sobre todos los valores de X, de $f(x)$ es uno haciendo uso del **Teorema del Binomio de Newton**:

$$(a+b)^n = \sum_{x=0}^{n} \binom{n}{x} a^{n-x} b^x,$$

en dónde si se reemplaza a por q y b por p se obtiene

$$(q+p)^n = \sum_{x=0}^{n} \binom{n}{x} q^{n-x} p^x = \sum_{x=0}^{n} f(x),$$

y como $q + p = (1-p) + p = 1$, se concluye que

$$\sum_{x=0}^{n} f(x) = (1)^n = 1.$$

Distribución Binomial

Su **Función Generatriz de Momentos** es:

$$M_X(t) = \mathrm{E}\left(e^{tX}\right) = \sum_{x=0}^{n} e^{tx} \cdot f(x) = \sum_{x=0}^{n} e^{tx} \cdot \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}$$

$$= \sum_{x=0}^{n} \left(\begin{array}{c} n \\ x \end{array} \right) (e^t)^x p^x (1-p)^{n-x} = \sum_{x=0}^{n} \left(\begin{array}{c} n \\ x \end{array} \right) (e^t p)^x (1-p)^{n-x}.$$

De acuerdo al binomio de Newton con $a = q, \; b = e^t p$:

$$M_X(t) = \mathrm{E}\left(e^{tX}\right) = (q + pe^t)^n.$$

Distribución Binomial

Media y Varianza

Utilizaremos la generatriz de momentos. Sabemos que $\mathrm{E}(X) = M'_X(0)$ y $\mathrm{E}(X^2) = M''_X(0)$, entonces

$$\begin{aligned} M'_X(t) &= \; n(q + pe^t)^{n-1}(pe^t), \\ M'_X(0) &= \; n(q + pe^0)^{n-1}(pe^0). \end{aligned}$$

Puesto que $e^0 = 1$ y $p + q = 1$, tenemos:

$$M'_X(0) = \; n(1)^{n-1}p = np,$$

y, por lo tanto,

$$\mathrm{E}(X) = np.$$

Distribución Binomial

Ahora,

$$
\begin{aligned}
M''_X(t) &= n(q+pe^t)^{n-1}pe^t + pe^t\left[n(n-1)(q+pe^t)^{n-2}pe^t\right], \\
\mathrm{E}(X^2) &= M''_X(0) = np + n(n-1)p^2.
\end{aligned}
$$

entonces

$$
\begin{aligned}
\mathrm{V}(X) &= \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = np + n(n-1)p^2 - (np)^2 \\
&= np + n^2p^2 - np^2 - n^2p^2 = np - np^2 = np(1-p) \\
&= npq.
\end{aligned}
$$

Distribución Binomial

La variable aleatoria Binomial como una suma de variables aleatorias Bernoulli

Hemos mencionado que cada X_i es una v.a. Bernoulli y que estas son independientes. Considera la variable aleatoria Y definida como la suma de n variables aleatorias Bernoulli(p), esto es, como la suma de ceros y unos producidos por los posibles resultados de las variables Bernoulli

$$Y = \sum_{i=1}^{n} X_i, \qquad \text{con } X_i \sim \text{Bernoulli}(p) \text{ e independientes entre sí.}$$

Podemos imaginarnos lo anterior como realizando n repeticiones de un experimento Bernoulli en el que cada resultado será independiente de los otros, un **muestreo aleatorio**.

Distribución Binomial

Recordando las propiedades del valor esperado y el valor que toma la esperanza de una v.a. *Bernoulli*(p) , tenemos que el valor esperado de Y es

$$\mathrm{E}(Y) = \mathrm{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{E}(X_i) = \sum_{i=1}^{n} p = np$$

y su varianza

$$\mathrm{V}(Y) = \mathrm{V}\left(\sum_{i=1}^{n} X_i\right).$$

Veremos posteriormente que la varianza de una suma de variables aleatorias **independientes** es la suma de las varianzas de cada v.a. individual (multiplicada por el cuadrado de su coeficiente respectivo), entonces

$$\mathrm{V}(Y) = \sum_{i=1}^{n} \mathrm{V}(X_i) = \sum_{i=1}^{n} pq = npq.$$

Distribución Binomial

Vemos que estos resultados coinciden con los de la v.a. Binomial(n, p) pero **no** es suficiente para afirmar que $Y \sim$ Binomial(n, p). Si su generatriz es la misma que la de una v.a. binomial podremos afirmar que Y es binomial:

$$
\begin{aligned}
M_Y(t) &= \mathrm{E}(e^{tY}) = \mathrm{E}\left(e^{t\sum_{i=1}^n X_i}\right) = \mathrm{E}\left(e^{t(X_1+X_2+\cdots+X_n)}\right) \\
&= \mathrm{E}\left(e^{tX_1}e^{tX_2}\ldots e^{tX_n}\right),
\end{aligned}
$$

de nuevo, como las variables aleatorias X_i son independientes, el último término se puede escribir como

$$
\begin{aligned}
M_Y(t) &= \mathrm{E}\left(e^{tX_1}\right)\mathrm{E}\left(e^{tX_2}\right)\cdots\mathrm{E}\left(e^{tX_n}\right) \\
&= (q+pe^t)(q+pe^t)\cdots(q+pe^t) = (q+pe^t)^n,
\end{aligned}
$$

que es idéntica a la generatriz de momentos de una binomial.

Distribución Binomial

Por lo tanto nuestra v.a. Y, producto de un muestreo determinado, es una v.a. Binomial(n, p).

¿Qué se obtiene si en la sumatoria que define Y usamos $n = 1$?

Es muy importante recalcar el hecho de que cada posible resultado del experimento es una v.a. con las mismas propiedades: misma distribución, mismo parámetro, y que el resultado de cada una de ellas será independiente del resultado de las otras.

Estas condiciones se resumen diciendo que se realizó un **Muestreo Aleatorio**.

Distribución Hipergeométrica

Si cambiamos la forma en que se realiza el muestreo y mantenemos como nuestra v.a. a

$$X = \# \text{ de "éxitos" en } n \text{ observaciones,}$$

nuestro modelo tiene que ser adaptado a las nuevas condiciones.

Supongamos que podemos hacer:

- n observaciones de un conjunto total de N posibles,
- la probabilidad de éxito en cada observación cambia "paso a paso".

Ahora estamos asumiendo que nuestra población es finita (N) y no hay independencia entre observaciones, esto es, el resultado de cada observación será afectado por los resultados anteriores.

Distribución Hipergeométrica

Como **ejemplo**, considera un lote de 25 "botes" de leche comercial de 1 litro, cada uno susceptible de ser clasificado como en mal estado (éxito = agria) o en buen estado. Se toma una muestra de tamaño n, se considera la posibilidad de que se tengan k botes con leche agria y se hace la revisión de cada uno de los botes. Aquí $X = \#$ de botes con leche en mal estado.

- Nuestra población inicial son los 25 "botes",
- La probabilidad de que un bote seleccionado contenga leche en mal estado cambiará conforme vamos haciendo el muestreo debido a que en el lote habrá inicialmente una cierta proporción de botes en mal estado (botes de leche en mal estado / N total de botes en el lote) y al ir sacando los botes esta proporción se verá afectada por los resultados obtenidos con anterioridad. Esto es, se realiza un muestreo sin reemplazo.

Distribución Hipergeométrica

Este tipo de experimento es muy usado cuando en la revisión se tiene que destruir o alterar el objeto, en nuestro caso hay que abrir el bote de leche para revisarlo.

Considera lo siguiente: si $k = 4$, la proporción inicial es de $\frac{4}{25}$, pero si sacamos un bote la proporción cambia a $\frac{3}{24}$ ó $\frac{4}{24}$, dependiendo de si el que se sacó contenía leche en mal estado o no, respectivamente.

Distribución Hipergeométrica

Para poder establecer el conjunto de posibles valores de X, necesitamos conocer, en principio, la proporción exacta de botes en mal estado, esto es, cuántos botes hay en mal estado en el lote de 25, y cuántas observaciones se hacen (n):

- Si hay $k = 3$ botes con leche agria y se toman $n = 5$ botes para revisarlos, X podrá tomar los valores $\{0, 1, 2, 3\}$, i.e. $0 \leq x \leq k$; pero si se revisan $n = 2$, X podrá tomar los valores $\{0, 1, 2\}$, i.e. $0 \leq x \leq n$.
- Algo similar ocurriría si el lote fuera de $N = 10$ botes, si hubiesen $k = 4$ botes con leche agria y si se revisaran $n = 7$: puesto que se están revisando 7 y sólo hay 6 en buen estado, X podría ser $\{1, 2, 3, 4\}$, i.e. $n - (N - k) \leq x \leq k$. ¿Qué valores toma X si $k = 5$ y $n = 7$?

Observa que x **llega hasta el menor de los números n y k, y que el menor valor de x es 0 ó $n - (N - k)$**. Estos casos muestran que los valores de la v.a. dependerán de los valores de k, n y N.

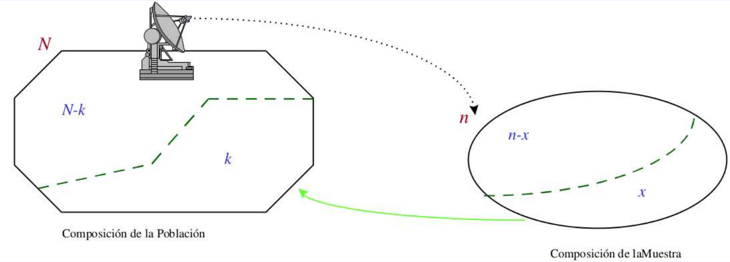
Distribución Hipergeométrica

Las condiciones de muestreo mencionadas suelen aplicarse en la práctica en problema de muestreo de **aceptación en control de calidad**, así como en la **estimación del tamaño de una población finita** N en **procesos de captura-recaptura**. (Ver [4] ejemplo 3.2.10 pág. 86, ejemplo 3.2.9 págs. 84-85. También te recomendamos (deberías leer) el capítulo 4, pág. 42 de [5], para muestreo de aceptación.)

No es de sorprender este tipo de aplicaciones ya que, como dijimos, al tomar n observaciones en las cuales habrá x "éxitos", conoceremos la proporción $\frac{x}{n}$, la cual nos dará información de la verdadera proporción $\frac{k}{n}$.

Distribución Hipergeométrica

Esquemáticamente podemos imaginarnos los parámetros de la distribución hipergeométrica como en la siguiente figura



Población con N objetos		Muestra de n objetos tomados al azar
k elementos con cierta característica de interés	$\leftarrow$ Inferencia	x elementos con la característica de interés
$N-k$ sin la característica	$\leftarrow$ Inferencia	$n-x$ sin la característica de interés

Distribución Hipergeométrica

La ley de comportamiento asociada con nuestra v.a. se obtiene utilizando el enfoque de frecuencia relativa en el que contamos el número de casos favorables y dividimos entre el número de casos totales:

- Nos interesan x "éxitos" de un total de k posibles, y esto lo podemos hacer de $\binom{k}{x}$ formas; pero por cada uno de estos resultados, como se seleccionaron n, obtendríamos $n-x$ "fracasos" de un total de $N-k$ "fracasos" posibles, y esto lo podemos hacer de $\binom{N-k}{n-x}$ formas. Entonces, usando la regla multiplicativa, el número total de **casos favorables** es $\binom{k}{x}\binom{N-k}{n-x}$.
- Al seleccionar n objetos de un total de N, podemos hacerlo de $\binom{N}{n}$ formas.

Distribución Hipergeométrica

Por lo tanto, la probabilidad de observar x "éxitos" en n pruebas está dada por la función de probabilidad (Distribución Hipergeométrica):

$$f(x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}, \qquad \text{máx}\{0, n-(N-k)\} \leq x \leq \text{mín}\{n, k\}.$$

Sus parámetros son n, k y N. Notación $X \sim \text{Hiper}(N, k, n)$.

Distribución Hipergeométrica

En las siguientes gráficas se muestra la funciones de masa y de distribución acumulada de la distribución hipergeométrica para varios valores de los parámetros.

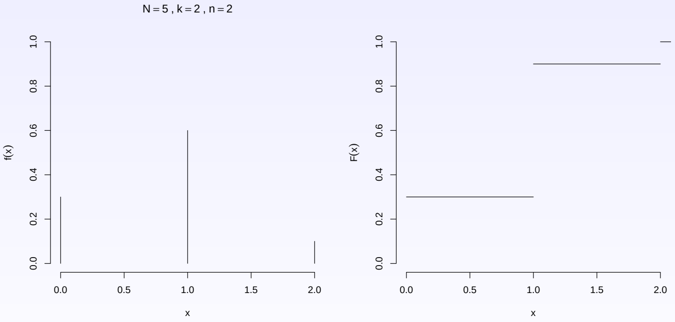


Figura: Funciones de masa y de distribución acumulada para $X \sim$Hiper$(5, 2, 2)$.

Distribución Hipergeométrica

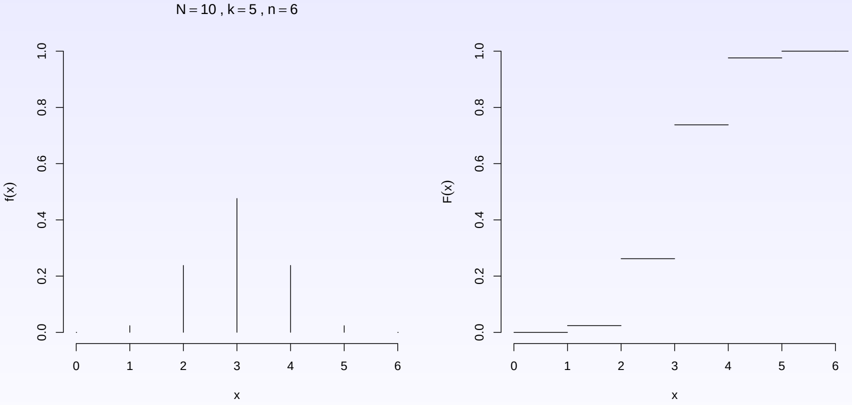


Figura: Funciones de masa y de distribución acumulada para $X \sim$Hiper$(10, 5, 6)$.

Distribución Hipergeométrica

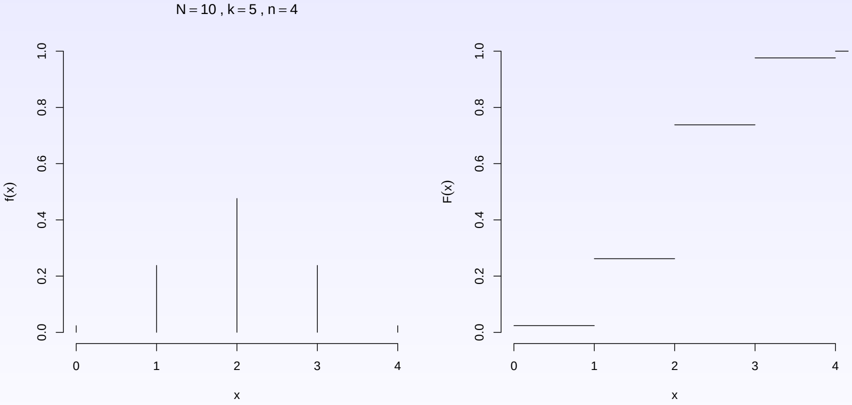


Figura: Funciones de masa y de distribución acumulada para $X \sim$Hiper$(10, 5, 4)$.

Distribución Hipergeométrica

Media y Varianza

La media y la varianza de una v.a. hipergeométrica están dados por:

$$\mathrm{E}(X) = \frac{nk}{N}, \qquad \mathrm{V}(X) = \frac{nk}{N}\left(1 - \frac{k}{N}\right)\left(\frac{N-n}{N-1}\right).$$

Debido a la especial complejidad y a su poco uso práctico, no se muestra la función generatriz de momentos de la hipergeométrica. (Ver referencias adicionales en Clase 2).

Distribución Hipergeométrica

Aproximación de la distribución binomial a la distribución hipergeométrica

Existe una relación entre la distribución binomial y la hipergeométrica: bajo ciertas condiciones, una puede aproximar los valores de probabilidad de la otra.

Observa que, a partir del enfoque de probabilidad como frecuencia relativa, la razón $\frac{k}{N}$ representa la proporción de éxitos p, mientras que $1 - \frac{k}{N}$ la proporción de fracasos q, de lo cual las anteriores fórmulas pueden ser escritas como

$$\mathrm{E}(X) = np, \qquad \mathrm{V}(X) = npq\left(\frac{N-n}{N-1}\right).$$

Distribución Hipergeométrica

Se puede apreciar una cierta similitud con la media y la varianza de la distribución binomial. Si N fuese muy grande, y el tamaño n de la muestra pequeño, tendríamos que

- el factor $\frac{N-n}{N-1}$ tendería a 1;
- p equivaldría a la probabilidad de "éxito" como en la distribución binomial, pudiéndose considerar como **"cuasi-fija"**, ya que no cambiaría mucho al ir realizando las observaciones;
- q equivaldría a la probabilidad de "fracaso", con las mismas consideraciones que para p; e,
- igualmente, podríamos considerar una **"cuasi-independencia"**, suponiendo que no tendríamos "rachas sistemáticas de resultados", esto es, el resultado de una observación no nos daría información del resultado de la siguiente observación.

Distribución Hipergeométrica

Con las condiciones anteriores nuestro modelo sería aproximadamente binomial, y podemos usar esta distribución como una aproximación al modelo hipergeométrico.

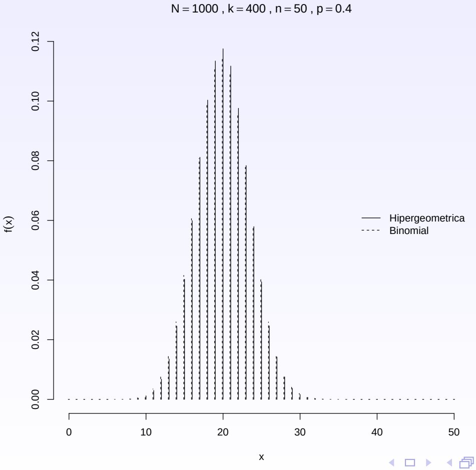
Es posible demostrar formalmente que si $p = k/N$ se mantiene constante entonces

$$\lim_{N\to\infty} \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} = \binom{n}{x} p^x (1-p)^{n-x}.$$

La pregunta es: ¿qué tan grande deberá ser N comparada con n? Se ha encontrado de la experiencia que cuando la proporción $\frac{n}{N}$ es del orden del 10 %, se tiene una buena aproximación, mejorando cuando $\frac{n}{N}$ disminuye.

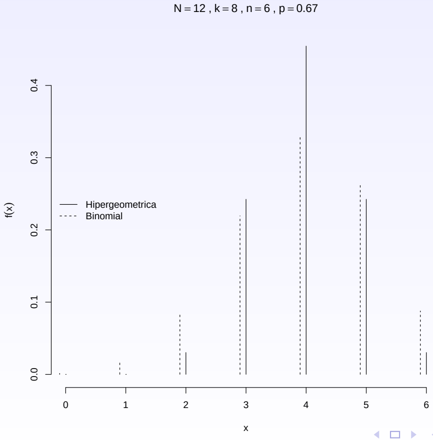
Distribución Hipergeométrica

Por ejemplo, tenemos que la aproximación binomial para $\text{Hiper}(1000, 400, 50)$ es excelente.



Distribución Hipergeométrica

En este segundo ejemplo la relación entre N y n es del 50 %, por lo que no se espera una buena aproximación entre las probabilidades asociadas bajo cada distribución.



Distribución Hipergeométrica

El factor $\frac{N-n}{N-1}$, cantidad que nos recuerda la finitud de N, es llamado el **factor de corrección para población finita**. ¿Cuál es $\lim_{N\to\infty} \frac{N-n}{N-1}$, con n fija?

Nota: Entre los materiales del curso se encuentra un texto sobre métodos de captura y recaptura asociados a la distribución hipergeométrica y sobre como calcular los momentos de tal distribución. Es importante leerlos.