UNIVERSIDAD DE MÁLAGA

Modelos probabilísticos

Variables aleatorias discretas

Carlos Gamero Burón
José Luis Iranzo Acosta
Departamento de Economía Aplicada
Universidad de Málaga

Parcialmente financiado a través del PIE13-024 (UMA)

Parcialmente financiado a través
de PIE13-024 (UMA)

Estadística I
Bloque II

GRADO EN ADMINISTRACIÓN Y DIRECCIÓN DE EMPRESAS

VARIABLE ALEATORIA Y MODELOS PROBABILÍSTICOS

Tema 4. PROBABILIDAD
Tema 5. VARIABLE ALEATORIA
Tema 6. MODELOS PROBABILÍSTICOS PARA VARIABLES ALEATORIAS DISCRETAS
Tema 7. MODELOS PROBABILÍSTICOS PARA VARIABLES ALEATORIAS CONTINUAS

Tema 6: MODELOS PROBABILÍSTICOS PARA VARIABLES ALEATORIAS DISCRETAS

6.1. Introducción

6.2. Distribución binomial

6.3. Distribución de Poisson

6.4. Distribución Hipergeométrica

6.5. Distribución multinomial

6.1. INTRODUCCIÓN

- Hasta ahora hemos introducido funciones matemáticas [$f(x)$ y $F(x)$] para representar distribuciones de probabilidad de variables aleatorias.

- Ahora estudiaremos las distribuciones de probabilidad más importantes (las de mayor uso) que sirven de modelos para explicar la realidad del experimento aleatorio que estamos estudiando.

- Un **modelo probabilístico** es una representación matemática deducida de un conjunto de supuestos con el doble propósito de estudiar los resultados de un experimento aleatorio y predecir su comportamiento futuro, cuando se realiza bajo las mismas condiciones dadas inicialmente.

- Dentro de los modelos de distribuciones discretas vamos a ver los siguientes:

  1. Distribución Binomial
  2. Distribución Poisson
  3. Distribución Hipergeométrica
  4. Distribución Multinomial (modelo multivariante)

6.2. DISTRIBUCIÓN BINOMIAL

- **Aplicaciones:** control de calidad, ventas, marketing, medicina, etc.

- **Experimento binomial:**

Ejemplo: $\xi =$"Arrojar una moneda *n* veces"

Características del experimento binomial:

Sólo hay dos resultados posibles en cada prueba (*prueba de Bernoulli*)
éxito = ocurrencia del suceso (*sacar cara*)
fracaso = no ocurrencia (*sacar cruz*).

El experimento se repite *n* veces (*n* pruebas) dándose que:

1. *n* es finito
2. las *n* pruebas son independientes entre sí.
3. *P*(éxito)=*p* se mantiene constante prueba a prueba
   P(fracaso)=1-*p*

- **Variable aleatoria:**

X="número de éxitos (caras) al realizar *n* lanzamientos"

X es una variable discreta: $x = 0,1,2,\ldots,n$

- **Función de cuantía:** $f(\mathrm{x}) = P(X = x)$

$$X \sim B(n,p) \quad \Leftrightarrow \quad f(x) = \begin{cases} \binom{n}{x} p^x q^{n-x} & x = 0,1,2,\ldots,n \\ 0 & \text{en el resto.} \end{cases}$$

Es una distribución biparamétrica (*n* y *p*)

Deducción de *f*(*x*):

$\xi =$"Arrojar una moneda *n* veces" es un experimento binomial

X="nº de caras que se obtienen"

"Cara"="éxito"=*E* y "Cruz"="fracaso"=*F*

El resultado global en *n* pruebas sucesivas, suponiendo que primero ocurren los *x* éxitos y después $n-x$ fracasos, puede escribirse así:

$$\underbrace{EE\ldots E}_{x}\,\underbrace{FF\ldots F}_{n-x}$$

La probabilidad de este suceso concreto sería:

$$P(EE\ldots EFF\ldots F)\underbrace{=}_{\substack{\text{por}\\ \text{independencia}}} P(E)P(E)\cdots P(E)P(F)P(F)\cdots P(F)=$$

$$=\underbrace{pp\cdots p}_{x\text{ veces}}\,\underbrace{qq\cdots q}_{n-x\text{ veces}}=p^x q^{n-x}$$

Da igual el orden en que aparezcan los *x* éxitos y los $n-x$ fracasos. Cualquier permutación con *x* éxitos y $n-x$ fracasos es válida:

$$P^n_{x,n-x}=\frac{n!}{x!(n-x)!}=\binom{n}{x}=k$$

Todas las ordenaciones H_i son equiprobables con $P(H_i)=p^x q^{n-x}$. La probabilidad de que ocurran *x* éxitos en *n* pruebas será la probabilidad de que ocurra cualquiera de esas ordenaciones:

$$P(X=x)=P(H_1\cup H_2\cup\cdots\cup H_k)\underbrace{=}_{\substack{\text{por ser}\\ \text{disjuntas}}}P(H_1)+P(H_2)+\cdots+P(H_k)=kp^x q^{n-x}$$

Luego la probabilidad buscada vendría dada por:

$$P(X=x)=\frac{n!}{x!(n-x)!}p^x q^{n-x}\qquad x=0,1,2,\ldots,n.$$

Función de distribución:

$$F(x) = P(X \le x) = \begin{cases} 0 & \text{si } x < 0 \\ \sum_{i=0}^{i=x} \binom{n}{i} p^i q^{n-i} & x = 0,1,2,\ldots,n-1 \\ 1 & \text{si } x \ge n \end{cases}$$

Podemos utilizar $F(x)$ para calcular la probabilidad de que X tome un valor concreto:

$$P(X = x) = F(x) - F(x-1).$$

Características de la distribución binomial:

- **Función generatriz de momentos:**

$$\boxed{M_X(t) = E\left[e^{tX}\right] = \sum_{x=0}^{x=n} e^{tx} f(x) = \left(pe^t + q\right)^n \quad \text{con } -\infty < t < +\infty.}$$

A partir de ella podemos obtener:

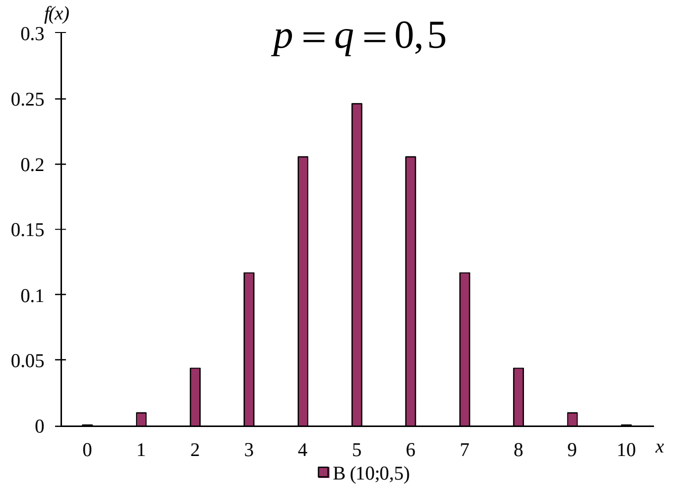
- **Esperanza:** $\boxed{E[X] = \mu = np.}$

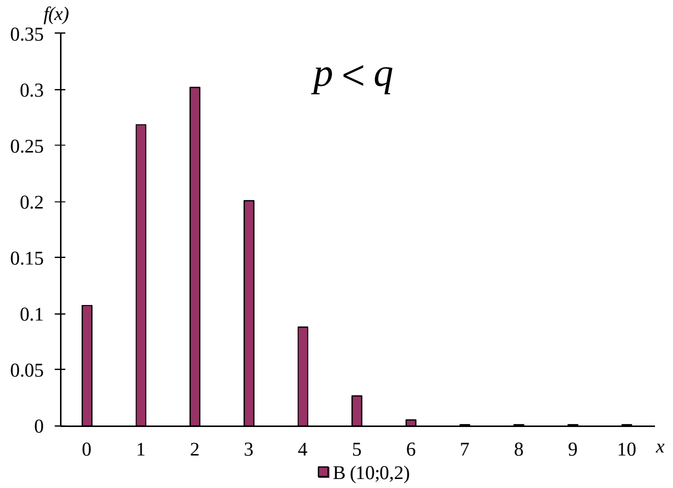
  El número medio de éxitos en n pruebas se obtiene multiplicando n por la probabilidad de éxito.

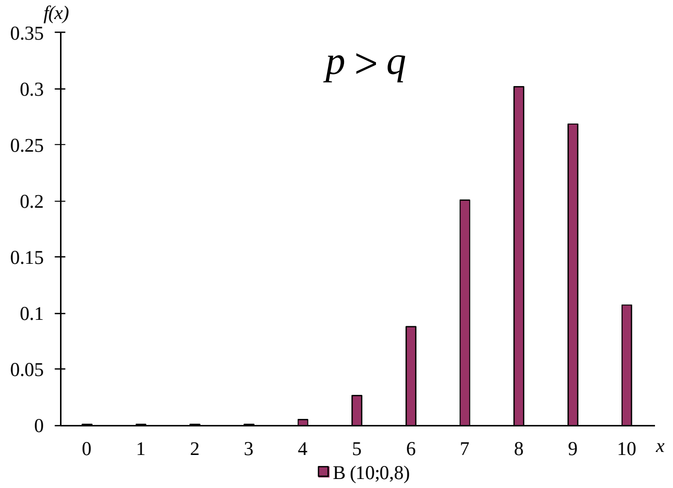
- **Varianza:** $\boxed{Var[X] = \sigma^2 = npq}$

μ y σ^2 dependen sólo de n y de p, los parámetros del modelo

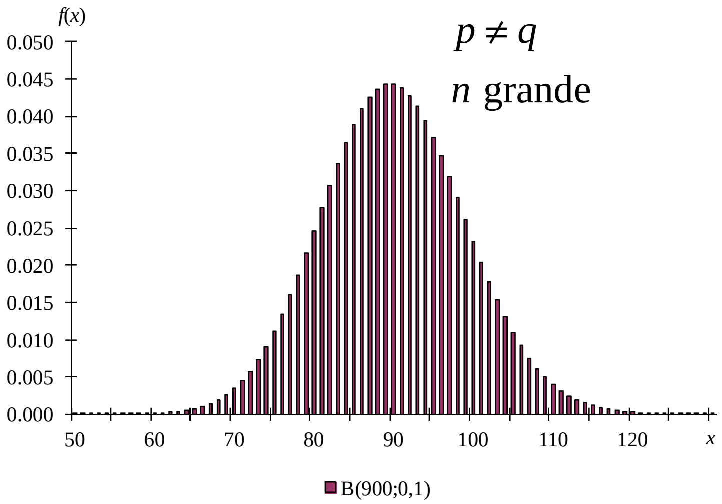
- **Forma de la distribución:**







Aunque p sea distinto de q, cuando n crece la distribución tiende a hacerse simétrica. Si n tiende a infinito, la distribución binomial tiende a aproximarse a una distribución normal



Cálculo exacto de probabilidades:

- Puede llevarse a cabo mediante el uso de programas estadísticos (Excel, SPSS, STATA, etc.).

- También existen tablas estadísticas que recogen los valores que toman la función de cuantía y la función de distribución para determinados valores de n y p (véase Anexo 5.1)

- Manejo de tablas:

  Importante:

  Sean $X \sim B(n,p)$ e $Y \sim B(n,1-p)$

  Relación entre funciones de cuantía:

$$P(X = x) = P(Y = n - x)$$

  Relación entre funciones de distribución:

$$F_X(x) = P_X(X \leq x) = P_Y(Y \geq n - x) = 1 - F_Y(n - x - 1)$$

Aproximación de probabilidades binomiales:

- Cuando n es grande, pueda usarse la distribución normal para aproximar probabilidades binomiales.

- Cuando n es grande, en determinadas situaciones, también se pueden aproximar utilizando la distribución de Poisson.

Propiedad reproductiva

Si X_1 y X_2 son dos variables aleatorias <u>independientemente</u> distribuidas según una $B(n_1, p)$ y $B(n_2, p)$, respectivamente (importante, con la misma probabilidad de éxito p), entonces:

$$X = X_1 + X_2 \sim B(n_1 + n_2, p).$$

Luego la distribución binomial se reproduce por adición de variables aleatorias binomiales independientes con la misma probabilidad de éxito p, respecto al parámetro n.

Demostración:

$$M_X(t) = E\left[e^{tX}\right] = E\left[e^{t(X_1+X_2)}\right] = E\left[e^{tX_1} \cdot e^{tX_2}\right] \underbrace{=}_{\substack{\text{por independecia} \\ \text{de } X_1 \text{ y } X_2}} E\left[e^{tX_1}\right] \cdot E\left[e^{tX_2}\right] =$$

$$= M_{X_1}(t) M_{X_2}(t) = \left(pe^t + q\right)^{n_1} \left(pe^t + q\right)^{n_2} = \left(pe^t + q\right)^{n_1+n_2}$$

Esta es la función generatriz de momentos de una $B(n_1 + n_2, p)$ por lo que, por el Teorema de la Unicidad, se concluye que $X = X_1 + X_2 \sim B(n_1 + n_2, p)$.

***Ejemplo 1*:**

Una máquina fabrica una determinada pieza y se sabe que un 7 por 1000 de las que produce son defectuosas. Hallar la probabilidad de que el examinar 50 piezas:

a) Sólo una sea defectuosa.

b) Se encuentren como máximo dos defectuosas.

Resolución:

$X =$ "Piezas defectuosas producidas" $\qquad X \sim B(50;0,007)$

a) $P\left(X=x\right)=f(x)=\binom{n}{x}p^x q^{n-x}$

$$f(1)=P\left(X=1\right)=\binom{50}{1}0,007^1 0,993^{49}=0,248$$

b) $P\left(X\le 2\right)=\binom{50}{0}0,007^0 0,993^{50}+\binom{50}{1}0,007^1 0,993^{49}+\binom{50}{2}0,007^2 0,993^{48}=0,99$

***Ejemplo 2*:**

Una Universidad estudia la posibilidad de comprar un lote grande de aparatos de detección de incendios. Por la experiencia de otras universidades se sabe que el 20% de estos aparatos suelen ser defectuosos. Si se seleccionan aleatoriamente 6 aparatos, a) ¿cuál es la probabilidad de que ninguno sea defectuoso?, b) ¿y de que haya más de uno defectuoso?

Resolución:

a) $X=$ "Nº de aparatos defectuosos de un total de $n=6$"

$p=P(\text{defectuoso})=0,2$

$$X \sim B(6;0,2) \text{ con } f(x)=\binom{6}{x}0,2^x 0,8^{6-x} \quad \text{para} \quad x=0,1,2,...,6$$

$$P(X=0)=\binom{6}{0}0,2^0 0,8^6=0,2621$$

b) $P(X>1)=1-\left[P(X=0)+P(X=1)\right]=1-\binom{6}{0}0,2^0 0,8^6-\binom{6}{1}0,2^1 0,8^5=$

$$\underbrace{=}_{\text{tablas}} 1-0,2621-0,3932=0,3446$$

Ejemplo 3:

El propietario de un establecimiento especializado en material informático estima que la probabilidad de que un cliente que entra en su establecimiento acabe comprando un equipo completo es de 0,2. ¿Cuántos clientes deben entrar para que la probabilidad de vender al menos un equipo completo sea de 0,9?

Resolución:

$X=$"Nº de clientes que compran un equipo completo de un total de n"
Éxito: cliente compra un equipo completo
Fracaso: cliente no compra un equipo completo
$P(\text{éxito})=p=0,2$
$P(\text{fracaso})=q=1-p=0,8$

Se tiene que $X\sim B(n;0,2)$ con $f(x)=\binom{n}{x}0,2^x 0,8^{n-x}$ para $x=0,1,2,\ldots,n$

Se pide que valor que tiene que tomar n para que $P(X\geq 1)=0,9$:

$$P(X\geq 1)=1-P(X<1)=1-P(X=0)=1-\binom{n}{0}0,2^0 0,8^n=0,9$$

$$1-0,8^n=0,9 \;\Rightarrow\; 0,8^n=0,1 \underbrace{\;\Rightarrow\;}_{\substack{\text{tomando}\\ \text{logaritmos}}} n\ln(0,8)=\ln(0,1) \;\Rightarrow\; n=10,32$$

El redondeo en el muestreo siempre se hace al alza, de manera que el número de clientes que deben entrar es de 11.

6.3. DISTRIBUCIÓN DE POISSON

Aplicaciones:

- Como aproximación al modelo binomial en determinadas circunstancias (lo veremos más adelante)
- Resolución de problemas que surgen al contar la frecuencia con que ocurren sucesos aleatorios, relativamente raros e independientes entre sí, en un intervalo fijo de tiempo, de espacio o similar (longitud, volumen, etc.) (*ley de los sucesos raros*)

  Ejemplos:

  - Número de accidentes que ocurren por unidad de tiempo (semana, día, etc.) en un cruce de carreteras o en una planta industrial
  - Número de llamadas telefónicas recibidas en una centralita por unidad de tiempo
  - Número de llegadas de aviones a un aeropuerto en una hora
  - Número de bacterias en un cm^3 de cierta preparación
  - Número de conejos por hectárea de monte.

  Éstos son fenómenos que no responden a una ley binomial. No tiene sentido hablar de fracasos.

Experimento aleatorio de Poisson:

1) El número de sucesos que ocurren en un intervalo de tiempo (región o volumen) específico es independiente del número de sucesos que ocurren en cualquier otro intervalo

2) La probabilidad de que un único suceso o resultado ocurra en un intervalo muy pequeño es proporcional a la longitud del intervalo (región o volumen) y no depende del número de sucesos que se produzca fuera del intervalo

3) La probabilidad de que ocurran más de un resultado o suceso en un intervalo (región o volumen) muy pequeño es prácticamente nula

Variable aleatoria:

X=”número de sucesos independientes que ocurren en el tiempo, espacio o volumen”

Función de cuantía:

$$X \sim P(\mu) \quad \Leftrightarrow \quad f(x)=\begin{cases} \dfrac{e^{-\mu}\mu^{x}}{x!} & x=0,1,2,\ldots \quad \text{con} \quad \mu>0 \\ 0 & \text{en el resto.} \end{cases}$$

IMPORTANTE: Ahora se considera que el número de eventos que pueden ocurrir puede ser infinito $(x=0,1,2,\ldots)$.

Esta distribución es uniparamétrica (μ): el parámetro μ caracteriza a la distribución y se corresponde con el **número medio de veces que ocurre el suceso aleatorio por unidad de tiempo, región o volumen.**

Función de distribución:

$$F(x)=P\left(X \le x\right)=\begin{cases} 0 & \text{si } x<0 \\ \displaystyle\sum_{i=0}^{i=x} \frac{e^{-\mu}\mu^{i}}{i!} & x=0,1,2,\ldots \end{cases}$$

Esta probabilidad se encuentra tabulada para distintos valores del parámetro μ.

Podemos obtener la probabilidad de que tome un valor concreto mediante la relación:

$$P\left(X=x\right)=F(x)-F(x-1).$$

Características:

- $M_X(t) = E\left[e^{tX}\right] = e^{\mu\left(e^t - 1\right)} \qquad -\infty < t < +\infty$

- **Esperanza:** $E[X] = \mu$

- **Varianza:** $Var(X) = \sigma^2 = \mu_2' - \mu^2 = \mu$

  La distribución de Poisson exhibe una importante propiedad: la media es igual a la varianza.

Ejemplo 1:

Se ha descubierto que el 13,5% de los discos duros portátiles vendidos por una empresa multinacional no contienen ningún sector defectuoso. Si suponemos que el número de sectores defectuosos por disco duro es una variable aleatoria con distribución de Poisson, determine el porcentaje de discos duros portátiles vendidos con un sector defectuoso.

Resolución:

$X =$ "Nº de sectores defectuosos por disco duro"

$Y \sim P(\mu)$ con $f(x) = \dfrac{e^{-\mu}\mu^x}{x!}$ para $x = 0, 1, 2, \ldots$

Desconocemos el valor de la media, pero el enunciado proporciona un dato que permite calcularla:

$$P(X=0) = f(0) = 0{,}135 \rightarrow \frac{e^{-\mu}\mu^0}{0!} = e^{-\mu} = 0{,}135 \underbrace{\rightarrow}_{\substack{\text{tomando}\\ \text{logaritmos}}} -\mu = \ln(0{,}135) \rightarrow \mu \simeq 2$$

El porcentaje solicitado es igual a

$$P(X=1)\cdot 100 = f(1)\cdot 100 = \frac{e^{-2}2^1}{1!}\cdot 100 = 27{,}07\%.$$

Propiedad reproductiva

Si X_1 y X_2 son dos variables aleatorias <u>independientes</u> y distribuidas según $X_1 \sim P(\lambda_1)$ y $X_2 \sim P(\lambda_2)$ entonces:

$$X = X_1 + X_2 \sim P(\lambda_1 + \lambda_2).$$

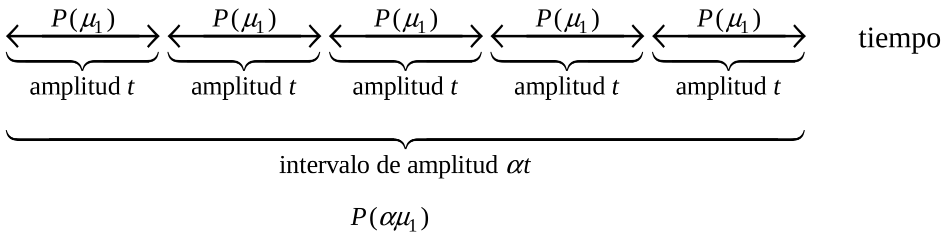
La suma de dos variables aleatorias independientes de Poisson es otra variable de Poisson cuyo parámetro es la suma de sus respectivos parámetros.

Demostración: a partir de funciones generatrices de momentos.

- Propiedad generalizable a la suma de *n* distribuciones de Poisson independientes.

- Se cumple la recíproca: si X_1 y X_2 son dos variables aleatorias independientes y la variable aleatoria $X = X_1 + X_2$ sigue una distribución de Poisson, entonces cada una de las variables X_1 y X_2 sigue una distribución de Poisson.

- En la práctica, estas propiedades llevan a si $X \sim P(\mu_1)$ donde μ_1 es el número medio de veces que aparece un suceso en un intervalo de amplitud *t*, entonces la distribución del número de sucesos que ocurren en otro intervalo de tiempo de amplitud αt sigue una distribución $P(\mu_2)$ donde $\mu_2 = \alpha\mu_1$:



La media de una distribución de Poisson es proporcional a la amplitud del intervalo considerado.

***Ejemplo 2*:**

El número de llamadas recibidas en un minuto en una línea telefónica sigue una distribución de Poisson con media igual a 4. Calcular:

a) Probabilidad de que no se reciba ninguna llamada en dos minutos.
b) Probabilidad de que no se reciba ninguna llamada en un intervalo de treinta segundos.

Resolución:

$X=$"número de llamadas recibidas en un minuto" $\quad X\sim P(4)$

$$f(x)=\begin{cases}\dfrac{e^{-4}4^{x}}{x!} & x=0,1,2,\ldots\\ 0 & \text{en el resto.}\end{cases}$$

Para resolver este ejercicio, aplicaremos la propiedad reproductiva:

a) $Y=$"número de llamadas recibidas en dos minutos"

En este caso, $\alpha=2$ de manera que $Y\sim P(2\cdot 4)$, es decir, $Y\sim P(8)$.

$$f(y)=\frac{e^{-8}8^{y}}{y!}\qquad y=0,1,2,\ldots$$

$$P\left(Y=0\right)=f(0)=\frac{e^{-8}8^{0}}{0!}=0,0003$$

b) $Z=$"número de llamadas recibidas en treinta segundos"

En este caso, $\alpha=\dfrac{1}{2}$ de manera que $Z\sim P\left(\dfrac{1}{2}\cdot 4\right)$, es decir, $Z\sim P(2)$.

$$f(z)=\frac{e^{-2}2^{y}}{z!}\qquad z=0,1,2,\ldots$$

$$P\left(Z=0\right)=f(0)=\frac{e^{-2}2^{0}}{0!}=0,1353.$$

Ejemplo 3:

Suponga que la llegada de vehículos a un cruce puede modelizarse por un proceso de Poisson. Si la probabilidad de que no pase ningún vehículo en un minuto es 0,135, se pide: ¿Cuál es la probabilidad de que pase más de un vehículo en dos minutos?

Resolución:

$X = \text{"Nº de vehículos que llegan a un cruce en 1 minuto"}$

$$X \sim P(\mu) \text{ con } f(x) = \frac{e^{-\mu}\mu^x}{x!} \text{ para } x = 0,1,2,\ldots$$

Desconocemos el valor de la media, pero el enunciado proporciona un dato que permite calcularla:

$$P(X=0) = 0{,}135 \rightarrow \frac{e^{-\mu}\mu^0}{0!} = e^{-\mu} = 0{,}135 \underbrace{\rightarrow}_{\substack{\text{tomando}\\\text{logaritmos}}} -\mu = \ln(0{,}135) \rightarrow \mu \simeq 2$$

$Y = \text{"Nº de vehículos que llegan a un cruce en 2 minutos"}$

La variable Y es la suma de dos variables aleatorias, X_1 y X_2, que se pueden suponer independientes y que se distribuyen según una Poisson $X_i \sim P(2)$. Por la propiedad reproductiva de la distribución de Poisson se tiene que:

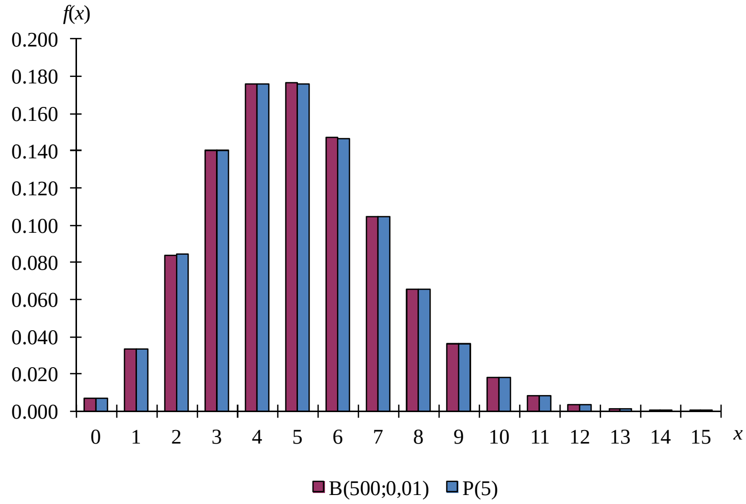
$$Y = X_1 + X_2 \sim P(2\mu) \quad \rightarrow \quad Y \sim P(4) \text{ con } f(y) = \frac{e^{-4}4^y}{y!} \text{ para } y = 0,1,2,\ldots$$

La probabilidad solicitada es:

$$P(Y>1) = 1 - P(Y \leq 1) = 1 - P(Y=0) - P(Y=1) \underbrace{=}_{\text{tablas}} 1 - \frac{e^{-4}4^0}{0!} - \frac{e^{-4}4^1}{1!} = 0{,}9084.$$

Aproximación de la distribución binomial a partir de la Poisson

- El <u>uso de una distribución para aproximar a otra</u> es una práctica bastante común en probabilidad y Estadística. La idea consiste en buscar situaciones en las que una distribución (como la de Poisson), cuyas probabilidades son relativamente fáciles de calcular, tiene valores que se encuentran razonablemente cercanos a las de otra distribución (como la binomial) cuyas probabilidades implican cálculos más complicados.
- La distribución de Poisson puede servir para obtener valores aproximados de las probabilidades binomiales, cuando n es muy grande, p muy pequeño y np permanece constante.
- En las aplicaciones prácticas se reemplaza habitualmente el modelo binomial por el de Poisson cuando, simultáneamente, $n \geq 50$ y $p \leq 0{,}1$.
- La aproximación es cada vez mejor a medida que n crece y p decrece.



Ejemplo 4:

Una máquina produce piezas en grandes cantidades y se sabe que la proporción de piezas defectuosas es *p*=0,01. Se toma una muestra aleatoria de 100 piezas. ¿Cuál es la probabilidad de que se obtengan 2 defectuosas?

$$X \equiv \text{"número de piezas defectuosas en una muestra de } n=100\text{"}$$

$$X \sim B(100;0,01)$$

$$P(X=2)=\binom{100}{2}0,01^2\cdot 0,99^{98}=0,1849$$

Utilizando la aproximación de Poisson con $\mu=100\cdot 0,01=1$:

$$P(X=2)\simeq\frac{e^{-1}1^2}{2!}=0,1839$$

El modelo de Poisson da una buena aproximación del modelo binomial.

Ejemplo 5:

Una compañía de seguros ha descubierto que alrededor del 0,1% de la población tiene cierto tipo de accidente cada año. Si los 10.000 asegurados fueran seleccionados aleatoriamente en la población, ¿cuál sería la probabilidad de que no más de 5 tengan este tipo de accidente en el próximo año?

$$X = \text{"número de accidentados en una muestra de } n=10000\text{"}$$

$$X \sim B(10000;0,001)$$

$$P(X\le 5)=\sum_{x=0}^{x=5}\binom{10000}{x}0,001^x\cdot 0,999^{1000-x}$$

Esta probabilidad resulta difícil de calcular. Podemos utilizar la aproximación de Poisson con $\mu=10000\cdot 0,001=10$:

$$P\left(X \leq 5\right) \simeq \sum_{x=0}^{x=5} \frac{e^{-10} 10^{x}}{x!} = F(5) \underbrace{=}_{\text{tablas}} 0,0671.$$

Ejemplo 6:

Una compañía de seguros ha realizado estudios que muestran que el 0,003% de los habitantes de una gran ciudad fallece cada año como resultado de un determinado tipo de accidente cubierto por sus pólizas. Calcule:

a) La probabilidad de que la compañía tenga que pagar a más de tres de los 10000 asegurados que tiene en esa ciudad.

b) El número esperado de accidentes del tipo mencionado entre sus asegurados.

Resolución:

$X =$ "Nº de asegurados que fallecen en un grupo de n=10000"

$P(\text{fallecimiento}) = p = 0,00003$

$X \sim B(10000; 0,00003)$

$$f(x) = \binom{10000}{x} 0,00003^{x} 0,00097^{10000-x} \quad \text{para} \quad x = 0,1,2,...,10000$$

a) $P(X > 3)$

Dada la problemática de cálculo, aproximaremos la distribución binomial por la distribución de Poisson. Se dan las condiciones para hacerlo ya que $n = 10000$ es suficientemente grande y $p = 0,00003$, suficientemente pequeña. La distribución de Poisson que utilizaremos para la aproximación será aquella que tiene la misma media que la binomial, es decir, $\mu = np = 10000 \cdot 0,00003 = 0,3$, de manera que $X_P \sim P(0,3)$:

$$P(X > 3) = 1 - P(X \leq 3) \simeq 1 - P(X_P \leq 3) = 1 - \sum_{x_P=0}^{x_P=3} \frac{e^{-0,3} (0,3)^{x_P}}{x_P\,!} \underbrace{=}_{\textit{tablas}} 0,0003$$

b) $E\left[X\right] = \mu = np = 0,3$ accidentes del tipo considerado mortales esperados.

6.4. DISTRIBUCIÓN HIPERGEOMÉTRICA

- **Aplicaciones:** control de calidad y aceptación de muestras.

- **Experimento aleatorio hipergeométrico:**

La distribución hipergeométrica es una alternativa a la distribución binomial cuando las **pruebas no son independientes** (**muestreo sin reposición**).

Cuando el muestreo o la selección de elementos de la muestra se hace sin reposición, la probabilidad de éxito no permanece constante como sucedía en el caso binomial.

Ejemplo:

Supongamos una caja con N=50 lámparas en la cual hay k defectuosas. Si realizamos el experimento de sacar una muestra de tamaño n=10 para comprobar si funcionan o no y el **muestreo es sin reposición**, **la probabilidad de que una lámpara sea defectuosa queda modificada por el resultado de las anteriores extracciones**

La distribución de probabilidad de la variable X que da el número de éxitos (número de elementos con una determinada característica) en la muestra de tamaño n obtenida sin reposición de una población de tamaño N es la distribución hipergeométrica.

- **Función de cuantía:**

Sea una población de tamaño N dividida en dos subpoblaciones disjuntas de tamaños k y $N-k$ (los que poseen la característica de interés y los que no).

Llamemos p a la proporción de elementos de la población que poseen la característica ($p=k/N$).

Se selecciona una muestra aleatoria sin reemplazamiento de tamaño n, con $n \leq N$.

X=”número de elementos con la característica de interés en una muestra de tamaño n extraída de una población de tamaño N”

La función de cuantía viene dada por:

$$f(x)=\begin{cases} \dfrac{\binom{Np}{x}\binom{N-Np}{n-x}}{\binom{N}{n}} & \max\left(0,n-(N-Np)\right)\le x\le \min(n,Np) \\ 0 & \text{en el resto.} \end{cases}$$

Ésta es la distribución hipergeométrica, que depende de tres parámetros N, n y p, con N y n siendo enteros positivos y $0<p<1$.

Notación abreviada: $X \sim H\left(N,n,p\right)$

La fórmula de la función de cuantía se obtiene como aplicación de la **Regla de Laplace**:

$\binom{N}{n}$= número de casos posibles. Es el número de muestras distintas de tamaño n obtenidas de la población total de tamaño N.

$\binom{Np}{x}$= número de formas posibles de obtener x elementos de la subpoblación que posee la característica de interés, que tiene Np elementos.

$\binom{N-Np}{n-x}$= número de formas posibles de obtener $n-x$ elementos de la segunda subpoblación, que tiene $N-Np$ elementos.

$\binom{Np}{x}\binom{N-Np}{n-x}$ = número de casos favorables. Es el número de formas distintas de obtener una muestra con x elementos de la primera subpoblación (la que posee la característica) y $n-x$ de la segunda.

La función de cuantía no es válida para $x=1,2,\ldots,n$, sino para $\max\left(0,n-(N-Np)\right)\leq x\leq \min(n,Np)$.

¿A qué se deben estas limitaciones en el valor de x?

- el número x de éxitos tiene que ser menor o igual que el número de elementos de la muestra (n), pero también tiene que ser menor o igual que el número de elementos con la característica de interés presentes en la población (Np). De ahí que $x\leq \min(n,Np)$.

- Por otro lado, el número de éxitos debe ser mayor o igual que cero, pero también el número de fracasos en la muestra debe ser menor o igual que el número de elementos de la población que no poseen la característica de interés $(n-x)\leq(N-Np)$. Por ello, $x\geq \max\left(0,n-(N-Np)\right)$.

Características de la distribución hipergeométrica:

- Media: $E(x)=np$

  La esperanza matemática es la misma que la de la distribución binomial

- Varianza: $\sigma^2=npq\dfrac{N-n}{N-1}$

La varianza del modelo hipergeométrico es menor que la del binomial ya que el factor $\dfrac{N-n}{N-1}$ es menor que la unidad. Este factor recibe el nombre de ***factor de corrección en muestreo con poblaciones finitas***. Representa la **reducción en la varianza por muestrear sin reposición en poblaciones finitas.**

Como este factor tiende a 1 cuando N tiende a infinito, se tiene que, para N suficientemente grande comparado con n, ambas distribuciones, la binomial y la hipergeométrica, tienden a hacerse idénticas por lo que, en ese caso, es poco relevante que la muestra de tamaño n sea con reposición o sin reposición. En la práctica se toma como límite $n < 0{,}1N \Rightarrow \frac{n}{N} < 0{,}1$.

Ejemplos:

(1) Se extrae con reposición una muestra de n=3 de una población de N=20. La probabilidad de éxito (tener una determinada característica) es 1/2. Tenemos pues que N=20, p=1/2, n=3. Al efectuarse el muestreo con reposición, la probabilidad de éxito se mantiene constante en cada extracción por lo que se trata de una distribución binomial.

(2) Imaginemos ahora que en esa misma población la extracción para la muestra se hace sin reposición (extracciones no independientes). Ahora los datos son N=20, p=1/2, Np=10, n=3. En esta situación, la probabilidad de éxito cambia de extracción a extracción: en la primera extracción es 10/20, en la segunda es 9/19 y en la tercera es 8/18. En este caso, la variable que recoge el número de éxitos sigue una distribución hipergeométrica.

(3) Pero si resultara que tenemos una población con N=1000 y Np=10, entonces, en el caso de extracción sin reposición, la probabilidad de éxito en la primera extracción sería 10/1000 y en la segunda 9/999, luego las diferencias son pequeñas (probabilidad casi constante). En este caso podríamos utilizar la distribución binomial en lugar de la hipergeométrica, sin cometer errores de gran magnitud.

***Ejemplo 1*:**

Un fabricante de automóviles compra los motores a una compañía donde se fabrican bajo estrictas especificaciones. El fabricante recibe un lote de 40 motores. Su plan para aceptar el lote consiste en seleccionar 8, de manera aleatoria y someterlos a prueba. Si encuentra que ninguno de los motores presenta serios defectos, el fabricante acepta el lote, de otra forma lo rechaza. Si el lote contiene dos motores con serios defectos ¿Cuál es la probabilidad de que sea aceptado?

Resolución:

$X =$ "nº de motores defectuosos con $n=8$, $N=40$ y $Np=2$"

El muestreo es sin reposición (no tiene sentido lo contrario) por lo que:

$X \sim H(40;2;0,05)$ con x=0,1,2

$$f(x)=\frac{\binom{Np}{x}\binom{N-Np}{n-x}}{\binom{N}{n}}=\frac{\binom{2}{x}\binom{38}{8-x}}{\binom{40}{8}} \quad x=0,1,2 \quad \text{y 0 en el resto.}$$

El lote es aceptado si no tiene ningún defecto:

$$f(0)=P(x=0)=\frac{\binom{2}{0}\binom{38}{8}}{\binom{40}{8}}=\frac{1\cdot\frac{38!}{8!\,30!}}{\frac{40!}{8!\,32!}}=\frac{992}{1560}=0,6359$$

Ejemplo 2:

Un fabricante asegura que sólo el 1% de su producción total se encuentra defectuosa. Supóngase que tiene 1000 artículos y se seleccionan 25 al azar para inspeccionarlos. Si el fabricante se encuentra en lo correcto. ¿Cuál es la probabilidad de observar dos o más artículos defectuosos en la muestra?

Resolución:

$X =$ "nº de artículos defectuosos con de $n=25$, extraida de una población con $N=1000$ y $p=0,01$"

Al no haber reposición, esta variable se distribuye según una hipergeométrica:

$X \sim H(1000;25;0,01)$ con x=0,1,2,…,10

$$f(x)=\frac{\binom{Np}{x}\binom{N-Np}{n-x}}{\binom{N}{n}}=\frac{\binom{10}{x}\binom{990}{25-x}}{\binom{1000}{25}} \quad x=0,1,2,\ldots,10 \quad \text{y 0 en el resto.}$$

La probabilidad solicitada es:

$$P(X\geq 2)=1-P(X<2)=1-\left[P(X=0)+P(X=1)\right]$$

$$P(x=0)=\frac{\binom{10}{0}\binom{990}{25}}{\binom{1000}{25}}=\frac{\dfrac{990!}{25!(900-25)!}}{\dfrac{1000!}{25!975!}}$$

$$P(x=1)=\frac{\binom{10}{1}\binom{990}{24}}{\binom{1000}{25}}=\frac{10\cdot\dfrac{990!}{24!(900-24)!}}{\dfrac{1000!}{25!975!}}$$

El cálculo de estas probabilidades resulta muy laborioso. Pero podemos aproximar sus valores mediante una binomial ya que $\frac{n}{N}<0{,}1$. Para este ejemplo utilizamos el modelo binomial $B(25;0,01)$:

$$f(x)=\binom{25}{x}0,01^{x}\cdot 0,99^{25-x}$$

$$f(0)=\binom{25}{0}0,01^{0}\cdot 0,99^{25}=0,7778 \qquad f(1)=\binom{25}{1}0,01^{1}\cdot 0,99^{24}=0,1964$$

$$P(X\geq 2)=1-P(X<2)=1-\left[P(X=0)+P(X=1)\right]=1-(0,7778-0,1964)=0,0258.$$

Ejemplo 3:

En un paquete de 100 memorias USB hay cuatro que contienen un bono de descuento para la compra de un *Netbook*. Calcule la probabilidad de que en una compra de nueve memorias seleccionadas al azar de dicho paquete se encuentren dos bonos. Compare este resultado con el que se obtendría utilizando alguna aproximación de la distribución exacta.

Resolución:

X = "Nº de memorias con bono en una muestra de tamaño $n = 9$ extraida de una población de tamaño $N = 100$ con $Np = 4$"

Tenemos una población formada por $N = 100$ memorias USB dividida en dos subpoblaciones, las que tienen bono descuento $(Np = 4)$ y las que no lo poseen $(N - Np = 96)$. La extracción de las $n = 9$ memorias se realiza *sin reposición*, de manera que estamos ante un experimento hipergeométrico

$$f(x) = P(X = x) = \begin{cases} \dfrac{\binom{4}{x}\binom{96}{9-x}}{\binom{100}{9}} & x = 0,1,2,3,4 \\ 0 & \text{en el resto.} \end{cases}$$

La probabilidad solicitada es:

$$P(X = 2) = f(2) = \frac{\binom{4}{2}\binom{96}{7}}{\binom{100}{9}} = \frac{\dfrac{4!}{2!2!}\cdot\dfrac{96!}{7!89!}}{\dfrac{100!}{9!91!}} = 0,0376$$

N es grande y $n / N < 0,1$, podríamos haber aproximado con la binomial:

$$\left.\begin{array}{l} p=\dfrac{4}{100}=0,04 \\ q=1-p=0,96 \\ n=9 \end{array}\right\} X_B \sim B(9;0,04) \;\rightarrow\; P(X_B=2)=\binom{9}{2}(0,04)^2(0,096)^7=0,0433.$$

Modelos de probabilidad para variables aleatorias discretas

Univariantes:

Binomial: $X \sim B(n,p)$ $\quad x=0,1,2,\ldots,n$

Poisson: $X \sim P(\mu)$ $\quad x=0,1,2,\ldots$

Hipergeométrico:

$$X \sim H(N,n,p) \qquad \max\big(0,n-(N-Np)\big)\le x\le \min(n,Np)$$

Multivariantes:

Multinomial: $X_1,\ X_2,\ldots,X_k \sim M\left(n,p_1,p_2,\ldots,p_k\right)$ $\quad \sum_{i=1}^{k} x_i=n$

6.5. DISTRIBUCIÓN MULTINOMIAL

- Este modelo probabilístico es un ejemplo de distribución discreta multivariante.
- Es una generalización de la distribución binomial. Recordemos que la distribución binomial permite resolver problemas de pruebas sucesivas independientes en experimentos dicotómicos, es decir, con dos resultados posibles en cada prueba, con probabilidad de éxito constante.

- **Experimento multinomial:**

Se pueden presentar k resultados mutuamente excluyentes $e_1, e_2, \ldots, e_k$ en cada prueba (recuérdese que en la binomial sólo se podían obtener dos resultados).

p_i es la probabilidad de que se presente e_i con $i = 1, 2, \ldots, k$ y $\sum_{i=1}^{k} p_i = 1$

Supongamos que se realizan n pruebas independientes en todas las cuales las probabilidades p_i permanecen constantes.

- **Variable aleatoria multidimensional:**

Consideramos k variables aleatorias $X_1, X_2, \ldots, X_k$, donde la variable X_i indica el número de veces que se presenta el resultado e_i cuando se realizan las n repeticiones independientes del experimento.

- **Función de cuantía multidimensional:**

$$f(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} \quad \text{con} \sum_{i=1}^{k} x_i = n \ \text{ y } \sum_{i=1}^{k} p_i = 1$$

Abreviadamente: $X_1, X_2, \ldots, X_k \sim M(n, p_1, p_2, \ldots, p_k)$

- **Función generatriz de momentos:**

$$M(t_1, t_2, \ldots, t_k) = \left(p_1 e^{t_1} + p_2 e^{t_2} + \cdots + p_k e^{t_k} \right)^n$$

A partir de la función generatriz de momentos es fácil comprobar que:

1. Las distribuciones marginales son binomiales: $X_i \sim B(n, p_i)$ con $i = 1, 2, \ldots, k$, lo que supone que $E[X_i] = np_i$ y $Var[X_i] = np_i(1 - p_i) = np_i q_i$.

2. Las variables $X_1, X_2, \ldots, X_k$ no son independientes dado que:

$$M(t_1, t_2, \ldots, t_k) \neq M_{X_1}(t_1) \cdot M_{X_2}(t_2) \cdots M_{X_k}(t_k)$$

***Ejemplo 1*:**

Supóngase que el 23% de los trabajadores de una gran empresa viven a menos de 5 kilómetros de su puesto de trabajo, el 59% de ellas viven entre 5 y 15 kilómetros de distancia y el 18% restante viven a más de 15 kilómetros. Se seleccionan al azar 20 empleados. Calcule la probabilidad de que siete de los seleccionados vivan a menos de 5 kilómetros, ocho vivan entre 5 y 15 kilómetros y cinco vivan a más de 15 kilómetros de su puesto de trabajo.

Resolución:

Comenzamos por identificar todos los elementos del problema:

n = 20 (número de trabajadores seleccionados),

k = 3 (número de grupos de clasificación de los trabajadores)

$X_1 = \{\text{número de trabajadores que viven a menos de 5 km.}\}$

$X_2 = \{\text{número de trabajadores que viven a una distancia de entre 5 y 15 km.}\}$

$X_3 = \{\text{número de trabajadores que viven a más de 15 km.}\}$

$p_1 = 0,23,\ p_2 = 0,59,\ p_3 = 0,18$

Teniendo en cuenta esta información, se tiene que:

$(X_1, X_2, X_3) \sim M(n, p_1, p_2, p_3) \rightarrow (X_1, X_2, X_3) \sim M(20; 0,23; 0,59; 0,18)$

Por tanto, la probabilidad solicitada es:

$$P\left(X_1=7, X_2=8, X_3=5\right)=\frac{20!}{7!8!5!}0,23^7\cdot 0,59^8\cdot 0,18^5=0,0094.$$

***Ejemplo 2*:**

Una agencia de demoscopia ha determinado que la probabilidad de que una persona vote a uno de los tres candidatos *A*, *B* y *C* para la alcaldía de una ciudad es, respectivamente, 0,1, 0,4 y 0,5. Suponiendo que en un programa de radio se realiza una encuesta a diez personas de esa ciudad elegidas al azar, se pide:

- **a)** Probabilidad de que el candidato *B* no obtenga ningún voto y los candidatos *A* y *C* el mismo número de votos.
- **b)** Probabilidad de que el candidato *A* obtenga los diez votos.
- **c)** Probabilidad de que el candidato *A* obtenga cinco votos.

Resolución:

Éste es un ejemplo de experimento multinomial con las siguientes características:

$$\left.\begin{array}{l} p_1=P(A)=0,1 \\ p_2=P(B)=0,4 \\ p_3=P(C)=0,5 \\ n=10 \end{array}\right\}(X_1,X_2,X_3)\sim M(10;0,1;0,4;0,5)$$

$$f(x_1,x_2,x_3)=\frac{10!}{x_1!x_2!x_3!}(0,1)^{x_1}(0,4)^{x_2}(0,5)^{x_3}\quad \text{para}\quad x_i=0,1,2,\ldots,10\quad \text{con}\ \sum_{i=1}^{i=3}x_i=10.$$

a) $P\left(X_1=5, X_2=0, X_3=5\right)=f(5,0,5)=\dfrac{10!}{5!0!5!}(0,1)^5(0,4)^0(0,5)^5=0,00008$

b) $P\left(X_1=10, X_2=0, X_3=0\right)=f(10,0,0)=\dfrac{10!}{10!0!0!}(0,1)^{10}(0,4)^0(0,5)^0=10^{-10}$

c) $P\left(A \text{ obtenga 5 votos}\right)$

La probabilidad solicitada es igual a:

$$P\left(A \text{ obtenga 5 votos}\right) = f(5,0,5) + f(5,5,0) + f(5,4,1) + f(5,1,4) + f(5,3,2) + f(5,2,3)$$

Resulta más fácil calcular esta probabilidad si tenemos en cuenta que las distribuciones marginales de las variables que intervienen en una distribución multinomial son distribuciones binomiales. Por tanto:

$$X_1 \sim B(10;0,1) \quad \rightarrow \quad P(X_1 = 5) = \binom{10}{5}(0,1)^5(0,9)^5 = 0,0015.$$