Probabilidad y Estadística (m) / Probabilidad (d)

Algunas distribuciones discretas famosas

Marina Valdora

Departamento de Matemática / Instituto de Cálculo

11 de abril de 2024

Outline

1. Algunas distribuciones discretas famosas
   - Bernoulli
   - Binomial
   - Hipergeométrica
   - Geométrica
   - Binomial Negativa
   - Poisson

Distribución Bernoulli

- Consideremos un experimento con dos posibles resultados: éxito y fracaso, es decir, $\Omega = \{E, F\}$. Se llama experimento o ensayo de Bernoulli.
- Sea X la variable aleatoria definida por $X(E) = 1$, $X(F) = 0$, es decir,

$$X = \begin{cases} 1 & \text{si el experimento resulta en un éxito} \\ 0 & \text{si el experimento resulta en un fracaso} \end{cases}$$

- Función de probabilidad puntual:

$$p_X(0) = 1 - p$$
$$p_X(1) = p$$

- Parámetro p: probabilidad de éxito.
- En este caso decimos que X es una variable aleatoria con distribución Bernoulli de parámetro p. Notación: $X \sim Be(p)$.

Outline

Distribución Binomial. Introducción con un ejemplo

Considere una urna con 5 bolillas blancas y 7 bolillas rojas. Se extraen 3 bolillas con reposición. Sea X el número de bolillas rojas extraídas.

1. Hallar el espacio muestral Ω y su cardinal.
2. Calcular $P(\omega)$, para cada $\omega \in \Omega$.
3. $\{X = k\} = \bigcup_{\omega \in \{X=k\}} \{\omega\}$. Calcular el cardinal de $\{X = k\}$. Deducir función de probabilidad puntual de X.

Distribución Binomial

- Consideremos el experimento: repetir n veces de manera independiente un ensayo Bernoulli.
- Sea X la cantidad de éxitos en las n repeticiones.
- Función de probabilidad puntual :

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Parámetros (n, p).
- Notación: $X \sim \mathcal{B}(n, p)$.

Deducción de la función de probabilidad puntual binomial

Repetimos un ensayo Bernoulli de parámetro p, n veces de forma independiente. Sea $X = \#$ de éxitos en n intentos.
Espacio muestral:

$$\Omega = \{\omega = (\omega_1, \ldots, \omega_n) : \omega_i \in \{0, 1\}\},$$

$\omega_i = 1$ si la i-ésima repetición fue un éxito.

1. Calcular $\#\Omega$.
2. Calcular $P(\omega)$, para cada $\omega \in \Omega$.
   $P(\omega) = \prod_{i=1}^{n} p^{\omega_i}(1-p)^{1-\omega_i} = p^{\sum_{i=1}^{n}\omega_i}(1-p)^{\sum_{i=1}^{n}(1-\omega_i)}$.
3. Observar que $P(\omega) = p^k(1-p)^{n-k}$ para todo $\omega \in \{X = k\}$.
4. $\{X = k\} = \bigcup_{\omega \in \{X=k\}} \{\omega\}$. Calcular el cardinal de $\{X = k\}$. Deducir función de probabilidad puntual de X.

Teorema 1

Consideremos el experimento: repetir n veces de manera independiente un ensayo Bernoulli. Sea X la cantidad de éxitos en las n repeticiones. Entonces

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Demostración.

$\square$

Ejemplo: Cálculo de Probabilidades

1. El 80 % de los alumnos de una cierta universidad son admitidos para continuar sus estudios en universidades estadounidenses. Si tres alumnos se postulan el próximo año, ¿cuál es la probabilidad de que exactamente dos de ellos sean admitidos?
2. Si 7 % de las piezas producidas por una fábrica presentan defectos, ¿cuál es la probabilidad de que al examinar 4 piezas sólo una de ellas esté fallada?
3. En cierta zona industrial, 7 de las 12 industrias existentes están en regla. 3 inspectores deciden visitar cada uno de ellos una de las empresas, en forma independiente (están peleados y no se hablan entre sí). Cada uno de ellos pasa su informe y considere X la cantidad de informes correspondientes a empresas en regla. Obtenga la función de probabilidad puntual de X.

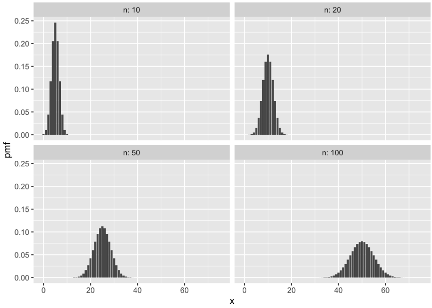
Ejemplo: Cálculo de Probabilidades

1.
2.
3.

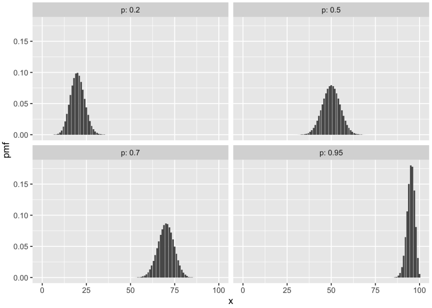
Ejemplo: Cálculo de Probabilidades en R

1. ```
   > dbinom(2, 3, 0.8)
   [1] 0.384
   ```
2. ```
   > dbinom(1, 4, 0.07)
   [1] 0.22522
   ```
3. ```
   > dbinom(0:3, 3, 7/12)
   [1] 0.07233796 0.30381944 0.42534722 0.19849537
   ```
   ```

Figura: Gráficos (de tipo histograma) de las funciones de probabilidad puntual de v.a. binomiales, para varios valores de n y p



(a) Distintos n y $p = 0{,}5$.



(b) $n = 100$ y varios p.

Outline

Distribución Hipergeométrica. Introducción con un ejemplo

Consideremos nuevamente una urna con 5 bolillas blancas y 7 bolillas rojas. Se extraen ahora 3 bolillas **sin reposición**. Sea X el número de bolillas rojas extraídas. Hallar el rango de la v.a. X. Hallar la función de probabilidad puntual de X.

Distribución Hipergeométrica

- Consideremos el experimento: extraer n elementos sin reemplazo de una población de tamaño N con M éxitos.
- Sea $X =$ cantidad de éxitos extraídos.
- $\Omega =$
- Función de probabilidad puntual:

$$p_X(k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

- $R_X = \{k \in \mathbb{N} : \max\{0, n - N + M\} \leq k \leq \min\{n, M\}\}$

Demostración.

$\square$

- Parámetros (N, M, n).
- Notación: $X \sim H(N, M, n)$.

Ejemplo: Cálculo de Probabilidades

1. En un lote de 30 piezas, se sabe que 5 tienen defectos. Se extraen 4 al azar. ¿Cuál es la probabilidad de que exactamente una de cuatro piezas esté defectuosa?
2. En una zona industrial, 7 de 12 industrias están en regla. Un inspectores visita tres empresas. Sea X la cantidad de empresas en regla entre las que visita. Calcular $p_X(k)$.

Si la muestra hubiese sido elegida con reemplazo, estaríamos en presencia de una variable aleatoria binomial con n repeticiones y $p = M/N$. Cuando el tamaño de la población es muy grande comparado con el tamaño de la muestra, podemos pensar que sacar con o sin reemplazo no debería alterar mucho la situación. Podemos entonces aproximar una hipergeométrica por una distribución Binomial obteniendo que

$$H(N, M, n) \text{``} \approx \text{''} \mathcal{B}(n, M/N) \ . \tag{1}$$

Para precisar el sentido de la igualdad aproximada dado en la ecuación (1) necesitamos la noción de convergencia en distribución, que introduciremos más adelante.

Ejemplo: Cálculo de Probabilidades

1. En un lote de 3000 piezas, se sabe que un 15 % tienen defectos. Se extraen 4 al azar. ¿Cuál es la probabilidad de que exactamente una de cuatro piezas esté defectuosa?

Hipergeométrica en R

En R, la función `dhyper` calcula la función de probabilidad puntual de la distribución Hipergeométrica:

- `dhyper(k, M, N-M, n)` calcula la $p_X(k)$ para $X \sim H(n, M, N)$.
- `phyper(k, M, N-M, n)` calcula la $F_X(k)$ para dicha distribución.

Hipergeométrica en R

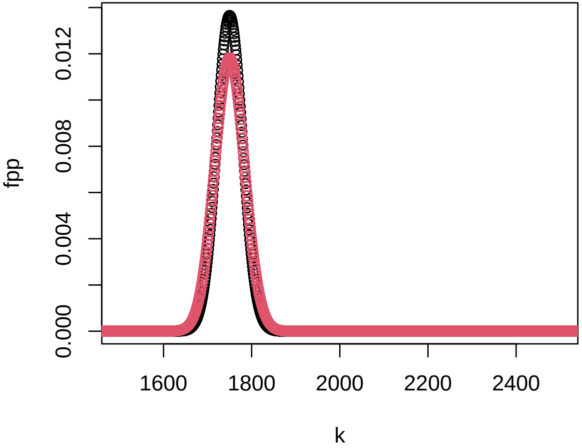
En un pueblo de 20000 ciudadanos se sabe que 7000 tienen un auto a su nombre y 13000 ninguno. Para un estudio de mercado se eligen 100 ciudadanos al azar ¿Cuál es la probabilidad de al menos que 40 de los 100 tengan un auto a su nombre?

```
> N = 20000
> M = 7000
> n = 100
> p = 7/20
> sum(dhyper(40:100, m=7000, n=13000, k=100))
[1] 0.1718243
> 1-phyper(39,m=7000, n=13000, k=100)
[1] 0.1718243
> # comparo con la binomial
> 1-pbinom(39, size = 100, prob = 7/20)
[1] 0.172415
```

Función de probabilidad puntual hipergeométrica y binomial

```
> ks = 0:5000
> plot(ks, dhyper(ks, M, N-M, 5000), xlim=c(1500, 2500),
ylab = "fpp", xlab = "k")
> points(ks, dbinom(ks, 5000, p), col = 2)
```

Función de probabilidad puntual hipergeométrica y binomial



Outline

Distribución Geométrica

- Consideremos el siguiente experimento: repetir un experimento de Bernoulli en forma independiente hasta que aparece un éxito.
- Función de probabilidad puntual:

$$p_X(k) = (1-p)^{k-1}p$$

- $R_X = \mathbb{N}$
- Parámetro p: probabilidad de éxito.
- Notación: $X \sim \mathcal{G}(p)$.

Propiedad de falta de memoria de la distribución geométrica

Proposición 1.1

Sea $X \sim \mathcal{G}(p)$, *para todo* $n, m \in \mathbb{N}$ *tenemos que*

$$P(X \geq n + m \mid X > m) \; = \; P(X \geq n).$$

La demostración es un ejercicio de la práctica.
Más aún, también es un ejercicio de la práctica probar que una variable aleatoria X con $R_X = \mathbb{N}$ posee la propiedad de falta de memoria discreta si y sólo si tiene distribución geométrica de parámetro $p = P(X = 1)$.

Otra forma de definir la Geométrica

- $X =$ cantidad de fracasos antes del primer éxito
- $R_X = \mathbb{N}_0$
- Función de probabilidad puntual:

$$p_X(k) = (1-p)^k p$$

- Si $X_1 =$ cantidad de repeticiones antes del primer éxito y $X_2 =$ cantidad de fracasos antes del primer éxito, entonces $X_1 = X_2 + 1$

Ejemplo: Cálculo de probabilidades

Un estudiante se presenta a un examen de ingreso a cierta universidad de tipo multiple choice. La probabilidad de que apruebe eligiendo las respuestas al azar es 0,1. ¿Cuál es la probabilidad de que consiga aprobar el examen en el sexto intento? Sea X el número de veces que es reprobado hasta aprobar. Hallar R_X y su función de probabilidad puntual.

Ejemplo: Cálculo de probabilidades en R

- Si ponen en R `help(dgeom)` van a ver que ese comando calcula la fpp de la geométrica y explica como define la v.a.:

  "the number of failures in a sequence of Bernoulli trials before success occurs."

```
> p <- 0.1
> p*(1-p)^5
[1] 0.059049
> dgeom(5, p)
[1] 0.059049
```

Outline

Distribución Binomial Negativa

- Consideremos el siguiente experimento: repetir un experimento de Bernoulli en forma independiente hasta que aparece el r-ésimo éxito en un ensayo Bernoulli.
- Definimos la variable aleatoria

$$X = \text{número de repeticiones hasta observar los } r \text{ éxitos.}$$

- $R_X = \{r, r+1, r+2, \cdots\}$.
- Función de probabilidad puntual:

$$p_X(k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}$$

- Parámetros (r, p).
- Notación: $X \sim \mathrm{NB}(r, p)$.

Otra forma de definir la binomial negativa

- $X =$ cantidad de fracasos antes del $r-$ésimo éxito
- $R_X = \mathbb{N}_0$
- Función de probabilidad puntual:

$$p_X(k) = \binom{k+r-1}{r-1} p^r (1-p)^k$$

Demostración.

$\square$

- Si $X_1 =$ cantidad de repeticiones antes del $r-$ésimo éxito y $X_2 =$ cantidad de fracasos antes del $r-$ésimo éxito, entonces $X_1 = X_2 + r$

Binomial negativa en R

The negative binomial distribution with `size` $= n$ and `prob` $= p$ has density

$$p(x) = \frac{\Gamma(x+n)}{\Gamma(n)x!} p^n (1-p)^x$$

for $x = 0, 1, 2, \ldots, n > 0$ and $0 < p \leq 1$.

This represents the number of failures which occur in a sequence of Bernoulli trials before a target number of successes is reached. The mean is $\mu = n(1-p)/p$ and variance $n(1-p)/p^2$.

Ejemplo con datos reales: Arbolado en espacios verdes de CABA en 2011

https://data.buenosaires.gob.ar/dataset/arbolado-espacios-verdes
Un biologo necesita para un experimento 4 arboles de más de 15 metros de altura. Planea elegir árboles al azar de la plaza Arenales y medirlos. ¿Cuál es la probabilidad de que los primeros 4 árboles que mida sean de más de 15 metros? ¿Cuál es la probabilidad de que necesite medir 6 para encontrar 4 árboles de más de 15 metros?

Ejemplo con datos reales: Arbolado en espacios verdes de CABA en 2011

```
> arbolado <- read.csv("arbolado-en-espacios-verdes.csv", sep=";")
> colnames(arbolado)
[1] "X"           "Y"           "ID_ARBOL"    "ALTURA_TOT"
[5] "DIAMETRO"    "INCLINACIO" "ID_ESPECIE" "NOMBRE_COM"
[9] "NOMBRE_CIE" "TIPO_FOLLA" "ESPACIO_VE" "UBICACION"
[13] "NOMBRE_FAM" "NOMBRE_GEN" "ORIGEN"      "COORD_X"
[17] "COORD_Y"
> arbolado[arbolado$ESPACIO_VE == "ARENALES",4]
[1]   5  5 28 22 23 18 18 14 20 16 32 23 17  5 36 25 27 28 12 16  4
[22] 19 19 25 32 14  1 16  4 22 21 15 17 16 23 22  6 25 38 32 13 15
[43] 24 32 35 22 33 28 11  5  5 16 25 23 29 22 19 19 16 18 32  2 19
[64] 10 11 24 13  6 12 12 16 19 20  9 16 27 10 32 20 23 20  5 17 32
[85] 21 17 18 32 15 10 32 25 11  2 15 14 20 27 17 10 21 22 13 16 28
[106] 20 18 15 16  5  4 18 12 16  5 28 17 22  8 23 19 24 21 20 22 2
[127] 22  5  4 20 25 20 23 16 14 20 16 11 14 14 11 28 19 18 20 18 1
[148] 19 17  4 15 20 18 28 22 13 11 27 30 12 20 12 12 18 32 16 14 3
[169] 28 12 13 23 34 19 17 22  4 11 24 20 22  5  5  5 26 11 25 12 2
[190] 22 12 22 16 22  9  6 12 15
```

Ejemplo con datos reales: Arbolado en espacios verdes de CABA en 2011

```
> arbolado_arenales = arbolado[arbolado$ESPACIO_VE == "ARENALES",]
> altura_arenales = arbolado_arenales$ALTURA_TOT
> p = mean(altura_arenales>15)

> dnbinom(0, size = 4, prob = p)
[1] 0.1801765
> dnbinom(2, size = 4, prob = p)
[1] 0.2188094
```

Generalización de la binomial negativa a $r \in \mathbb{R}$

Recordemos la función Gamma, $\Gamma : \mathbb{R}_{>0} \to \mathbb{R}$ dada por

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx.$$

Esta función definida mediante la fórmula precedente está bien definida (como integral impropia) y goza de las siguientes propiedades:
1) Integrando por partes, obtenemos que $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$.
2) Siendo $\Gamma(1) = 1$, de la propiedad anterior tenemos que $\Gamma(n) = (n - 1)\,!$ para todo $n \in \mathbb{N}$. Es por ello que decimos que la funcion Γ extiende el concepto de factorial para números positivos.
3) Haciendo el cambio de variables $x = \frac{u^2}{2}$, puede mostrarse que $\Gamma(1/2) = \sqrt{\pi}$.

Generalización de la binomial negativa a $r \in \mathbb{R}$

Función de probabilidad puntual:

$$p_X(k) = \frac{\Gamma(k+r)}{\Gamma(r)k!} p^r (1-p)^k$$

Observación: Se puede probar que, aún cuando r no es entero, esta es una función de probabilidad puntual bien definida.

Outline

Distribución de Poisson

Definición 2

Sea X una v.a. con función de probabilidad puntual:

$$p_X(k) = \frac{e^{-\lambda}\lambda^k}{k!}, \text{ para } k \in \mathbb{N}_0$$

entonces decimos que X tiene distribución de Poisson con parámetro λ.

Observación 1.1

La función p_X definida en el item anterior cumple las propiedades de una función de probabilidad puntual.

Demostración.

$\square$

- Notación: $X \sim \text{Pois}(\lambda)$.

Poisson

Lema 3

Sea $\lambda > 0$ *y* $(p_n)_{n\geq 1}$ *una sucesión de números en* $[0,1]$ *tal que* $np_n \to \lambda$. *Sea* $Y_n \sim \mathcal{B}(n, p_n)$, *entonces*

$$\lim_{n\to\infty} p_{Y_n}(k) = p_X(k), \quad \textit{para todo } k \in \mathbb{N}_0,$$

con $X \sim \mathcal{P}(\lambda)$.

Poisson

Demostración.

Sabemos que $np_n \to \lambda$. Esto implica que $p_n \to 0$.

$$p_{Y_n}(k) = \binom{n}{k} p_n^k (1-p_n)^{n-k} = \frac{\overbrace{n(n-1)\cdots(n-k+1)}^{k \text{ factores}}}{k!} p_n^k (1-p_n)^n (1-p_n)^{-k}$$

$$\lim_{n\to\infty} (1-p_n)^n = \lim_{n\to\infty} \left[(1+(-p_n))^{-\frac{1}{p_n}}\right]^{-np_n} = e^{-\lambda}$$

$$\lim_{n\to\infty} n(n-1)\cdots(n-k+1)p_n^k = \lim_{n\to\infty} \prod_{i=0}^{k-1} [(n-i)p_n] = \lambda^k$$

$$\lim_{n\to\infty} (1-p_n)^{-k} = 1$$

Juntando todo tenemos

$$\lim_{n\to\infty} p_{Y_n}(k) = \lambda^k e^{-\lambda} \frac{1}{k!} = p_X(k), \quad X \sim \mathcal{P}(\lambda)$$

Distribución de Poisson. Aplicación

- La distribución de Poisson suele emplearse para modelar la cantidad de éxitos que ocurren en una situación que puede pensarse que hay una cantidad muy grande de ensayos Bernoulli y la probabilidad de éxito en cada uno es muy pequeña.
- Por ejemplo, sea

  $$X = \text{ cantidad de incendios que ocurren en una cierta ciudad en un año}$$

  Imaginemos que podemos sudividir el tiempo en n subintervalos $(t_i; t_i + \Delta n]$ tan cortos como para que en cada subintervalo pueda ocurrir a lo sumo un éxito (un incendio). Si además suponemos que, en todos los intervalitos, la probabilidad de que ocurra un incendio se mantiene constante y que lo que ocurre en un intervalito es independiente de lo que ocurre en otro, entonces la cantidad de incendios en un año seguiría una distribución binomial.
- Esto sugiere que podemos modelar la distribución de X como el límite de distribuciones binomiales, es decir, con una distribución Poisson.

Distribución de Poisson. Más ejemplos de aplicación

- La cantidad de reclamos de un tipo (por ejemplo, robo de hogares) que recibe una compañia de seguros durante un año.
- La cantidad de errores de tipeo en una página de texto publicada.
- La cantidad de partículas emitidas por un determinado volumen de cierto material radiactivo en una hora.

La Poisson en R

En R, la función `dpois` calcula la función de probabilidad puntual de la distribución de Poisson y la función de distribución es `ppois`.

Distribución de Poisson. Ejemplo

La cantidad de errores de tipeo en una página de texto publicada por cierta editorial sigue una distribución de Poisson de parámetro 0,1. Hallar la probabilidad de encontrar más de un error de tipeo en una página.

En R:

```
 1 - dpois(0, 0.1) - dpois(1,0.1)
 [1] 0.00467884
```

Relación entre la binomial negativa y la Poisson

Proposición 1.2

Sea $\lambda > 0$ *y sea*

$$Y_n \sim \mathcal{BN}\left(n, \frac{n}{n+\lambda}\right),$$

entonces $\lim_{n\to\infty} p_{Y_n}(k) = p_X(k)$, *para todo* $k \in \mathbb{N}_0$, *con* $X \sim \mathcal{P}(\lambda)$.

Demostración.

$\square$