Curso Propedéutico de Probabilidad y Estadística

Gudelia Figueroa P.

Conceptos básicos de Probabilidad

Un *conjunto* es una colección de elementos. Denotamos por

$$A = \{\xi_1, \xi_2, \ldots, \xi_n\}$$

a un conjunto donde ξ_i es el $i-$é*simo* elemento en el conjunto.

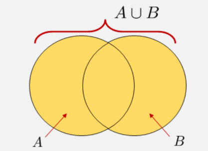
Para decir que un elemento ξ pertenece al conjunto A, lo denotaremos por $\xi \in A$. Por otra parte, B es un *subconjunto* de A, si para cualquier $\xi \in B$, ξ también está en A, lo cual se denota como $B \subseteq A$.

Un conjunto es *vacío* si no contiene elementos, y se denota como $A = \emptyset$.

Conceptos básicos

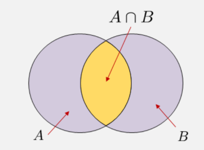
- **Unión.** La unión de dos conjuntos A y B contiene todos los elementos que están en A ó en B. Esto es,

  $$A \cup B = \{\xi | \xi \in A \quad ó \quad \xi \in B\}.$$

- **Intersección.** La intersección de dos conjuntos A y B contiene todos los elementos en A y en B. Esto es,

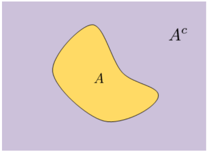
  $$A \cap B = \{\xi | \xi \in A \quad y \quad \xi \in B\}.$$

Conceptos básicos

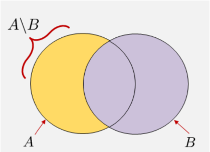
- **Complemento.** El *complemento* de un conjunto A es el conjunto que contiene todos los elementos que están en Ω pero no en A y se denota por A^c.

$$A^c = \{\xi | \xi \in \Omega \quad and \quad \xi \notin A\}.$$



Diferencia. La *diferencia* $A \setminus B$ es el conjunto que contiene todos los elementos que están en A pero no en B.

$$A \setminus B = \{\xi | \xi \in A \quad y \quad \xi \notin B\}.$$



Conceptos básicos

- **Conjuntos disjuntos.** Dos conjuntos A y B son *disjuntos* si $A \cap B = \emptyset$.
- En el caso de una colección de conjuntos $\{A_1, A_2, \ldots, A_n\}$, se dice que son disjuntos si, para cualquier par $i \neq j$, $A_i \cap A_j = \emptyset$.

Probabilidad

*Experimento aleatorio.*Es todo aquel experimento que cuando se repite bajo las mismas condiciones iniciales, el resultado no siempre es el mismo.

- Lanzar una moneda.
- Lanzar un dado

Probabilidad

Tres elementos constituyen un *espacio de probabilidad*:

- *Espacio muestral* Ω. El conjunto de todos los posibles resultados de un experimento.
- *Espacio de eventos* $\mathcal{F}$. La colección de todos los posibles eventos. Un evento E es un subconjunto en Ω que define un resultado o una combinación de resultados.
- *Ley de probabilidad* P. Un mapeo de un evento E a un número $P(E)$ que mide el tamaño del evento.

Probabilidad

Una *ley de probabilidad* es una función $P : \mathcal{F} \to [0,1]$ que mapea un evento A a un número real en $[0,1]$. Esta función debe satisfacer los axiomas de probabilidad:

1. $P(A) \geq 0$, para cualquier $A \in \mathcal{F}$
2. $P(\Omega) = 1$.
3. Para cualesquiera conjuntos disjuntos $\{A_1, A_2, \ldots, A_n\}$, debe ser cierto que: $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

Probabilidad

Denotemos por $P(A)$ la probabilidad de ocurrencia de un evento A.

- $0 \leq P(A) \leq 1$
- $P(A^c) = 1 - P(A)$
- $P(\emptyset) = 0$, donde, A^c representa el *complemento* de A, y $P(A^c)$ puede interpretarse como la probabilidad de que el evento A no ocurra.

Reglas de la probabilidad

Consideremos dos eventos A y B. Definimos la unión de estos eventos, como:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Luego entonces, podemos definir la probabilidad condicional de A dado B, como

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)},$$

y entonces la intersección de A y B eventos se calcula como:

$$P(A \cap B) = P(A \mid B)P(B).$$

Variable aleatoria

Matemáticamente, una variable aleatoria X, en un espacio muestral Ω, es una función de Ω a los números reales. Informalmente, puede decirse que una variable aleatoria es una variable cuyo resultado numérico depende de un experimento aleatorio.

Definición. Una *variable aleatoria* X es una función $X : \Omega \to \mathbb{R}$ que mapea un resultado $\xi \in \Omega$ a un número $X(\xi)$ en la recta real.

Por lo general utilizaremos letras mayúsculas $X_1, X_2, ..., X_n$ para denotar variables aleatorias y letras minúsculas $x_1, x_2, ..., x_n$ para sus correspondientes valores.

Variable aleatoria discreta

Una variable aleatoria X es discreta si su conjunto de valores posibles es finito o infinito numerable.

Una variable aleatoria discreta X toma valores $x_1, x_2, \ldots, x_n$, con probabilidad $P(x_1), \ldots, P(x_n)$, donde $\sum_{i=1}^{n} P(x_i) = 1$. Estas probabilidades reciben el nombre de función de masa o función de probabilidad.

Definición. La *función de probabilidad* de una variable aleatoria X es una función que especifica la probabilidad de obtener un número $X(\xi) = x$ y puede denotarse como

$$p_X(x) = P(X = x).$$

La función de masa o función de probabilidad puede ser visualizada como un diagrama de barras donde se visualiza la probabilidad de cada uno de los valores que X toma.

Ejemplos de variables aleatorias discretas

Ejemplo: Consideremos el lanzamiento de una moneda legal, y X denota el número de caras que aparecen en 20 lanzamientos de la moneda. Esto es, X puede tomar los valores 0 a 20, cada uno con probabilidad 1/2.

¿Otros ejemplos de variables aleatorias discretas?

Variable aleatoria continua

Una variable aleatoria continua X está definida por su función de densidad de probabilidad f_X.

Definición. Una función de densidad de probabilidad f_X de una variable aleatoria X, es un mapeo $f_X : \Omega \to \mathbb{R}$ que, cuando se integra sobre un intervalo $[a, b]$, proporciona la probabilidad de que $a \leq X \leq b$:

$P(a \leq X \leq b) = \int_a^b f_X(x)dx.$

Si la variable aleatoria X puede tomar valores en un continuo de valores, digamos cualquier valor en un intervalo a and b, entonces no es posible listar los valores x_i que toma la variable, ni asignarles una probabilidad p_i para $X = x_i$; pues para cualesquier valor x_i, $P(X = x_i)$ es cero.

¿Ejemplos de variables aleatorias continuas?

Función de distribución acumulada

Podemos, definir la *función de distribución acumulada*

$$F(x) = P(X \leq x),$$

En el caso de una variable aleatoria discreta,

$$F(x) = \sum_k P(X \leq x_k).$$

donde se suma sobre todos los índices k que satisfacen $x_k \leq x$.
Cuando se tiene una variable aleatoria continua

$$F(x) = \int_{-\infty}^{x} f(t)\, dt.$$

Propiedades de la función de distribución acumulada

- $0 \leq F(x) \leq 1$, para toda x
- $\lim_{x \to -\infty} F(x) = 0$
- $\lim_{x \to \infty} F(x) = 1$
- $F(x)$ es una función no-decreciente de x

Valor esperado de una variable aleatoria discreta

Definición. Sea X una variable aleatoria discreta con valores posibles $x_1, \ldots, x_n \ldots$. Sea $p_X(x_i) = P(X = x_i)$, $i = 1, 2, \ldots,$ El *valor esperado* de X denotado por $E(X)$ se define como

$$E(X) = \sum_{i=1}^{\infty} x_i \, p_X(x_i)$$

- Siempre que $\sum_{i=1}^{\infty} x_i \, p_X(x_i) < \infty$. Si la suma diverge, el valor esperado no existe.

En el caso del ejemplo de lanzamiento de un dado, tendremos:

$$E(X) = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = \frac{7}{2}.$$

Valor esperado de una variable aleatoria continua

El valor esperado de una variable aleatoria X, continua, es:

$$E(X) = \int_{-\infty}^{\infty} x\, f_X(x)\, dx$$

- Siempre que $\int_{-\infty}^{\infty} |x|\, f_X(x)\, dx < \infty$. Cuando la integral diverge, el valor esperado no existe.

Revisar propiedades del valor esperado.

Varianza de una variable aleatoria

La varianza de una variable aleatoria X se define como

$$V(X) = E[X - E(X)]^2, \tag{1}$$

la cual también podemos calcular como:

$$V(X) = E(X^2) - [E(X)]^2 \tag{2}$$

Revisar propiedades de la varianza.

Algunas variables aleatorias discretas

Variable aleatoria Bernoulli

La notación $X \sim \text{Bernoulli}(p)$ denota una variable aleatoria X que sigue una distribución Bernoulli, con parámetro p, donde $0 < p < 1$.

Una variable aleatoria X Bernoulli, con parámetro p, tiene una función de probabilidad dada por

$$P(X = x) = p^x (1 - p)^{1-x} \qquad\qquad x = 0, 1$$

para $0 < p < 1$.

Esta distribución está asociada a la noción de ensayo de Bernoulli, que es un experimento aleatorio con dos posibles resultados, que son agrupados en dos conjuntos excluyentes referidos como $\{\text{éxito}\}$ $(x = 1)$ y $\{\text{fracaso}\}$ $(x = 0)$.

Ejemplos de experimentos con v.a. Bernoulli

- Lanzamiento de una moneda, donde $E = \{cara\}$ y $F = \{cruz\}$.
- Lanzamiento de un dado, donde $E = \{2, 4, 6\}$ y $F = \{1, 3, 5\}$.
- Al estudiar la duración de llamadas: $E =$\{Duración ≥ 3 mins\} y $F =$ \{Duración < 3 mins\}.
- **¿Otros ejemplos?**

Bernoulli: valor esperado y varianza

$$E[X] = 0 * (1 - p) + 1 * p = p$$

$$V[X] = E[X^2] - E[X]^2 = 0^2 * (1 - p) + 1^2 * p - p^2 = p(1 - p)$$

Distribución Binomial

Considere n ensayos independientes de Bernoulli, cuyo resultado puede ser with "éxito" o "fracaso". Cada uno de estos ensayos tiene P(éxito)=p y P(fracaso)=$1-p$. La variable aleatoria que denota:
X : Número de éxitos en n ensayos independientes, es una variable aleatoria binomial, con parámetros n y p. Esto es,

Si $X \sim Bin(n,p)$, entonces

$P_X(x) = P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$

Supuestos distribución binomial

El modelo binomial asume:

- Número fijo de ensayos.
- Cada ensayo tiene dos posibles resultados.
- Probabilidad de éxito permanece constante de ensayo a ensayo.
- Los ensayos son independientes.

Ejemplo

Suponga que un 40% de la población de mexicanos tienen el esquema completo de vacunación contra Covid-19. Si se selecciona aleatoriamente a 10 habitantes de esta población,

- ¿cuál es la probabilidad de que resulten sólo tres con el esquema completo?
- ¿Cuál es la probabilidad de que a lo más 4 tengan el esquema completo?
- ¿Cuál es la probabilidad de que 8 o más tengan el esquema completo ?

Esperanza de una variable aleatoria

Hemos definido la esperanza de una variable aleatoria discreta como:

$$E(X) = \sum_{x} x\, P(X = x)$$

Para el caso de una variable aleatoria X binomial, ¿Cuál es su valor esperado?

$E(X) = \sum_{x=0}^{n} x \binom{n}{x} p^{x}(1-p)^{n-x}\}$. Veamos que desarrollando lo anterior podemos llegar a que $E(X) = np$:

Varianza de una variable aleatoria

La varianza de una variable aleatoria discreta la podemos calcular como:

$V(X) = E(X^2) - [E(X)]^2$,

Para calcularlo necesitamos conocer:

$$E(X^2) = \sum_x x^2 \, p_X(x)$$

En el caso de una variable aleatoria binomial es más sencillo calcular $E(X(X-1)) = n(n-1)p^2$.

De esta manera
$V(X) = E(X^2) - [E(X)]^2 = E(X(X-1)) + E(X) - E(X)^2 =$
$n(n-1)p^2 + np - (np)^2 = np(1-p)$

Distribución Binomial: Ejemplo

Si X es una variable aleatoria binomial con parámetros $n = 7$ y $p = 1/4$, calcular la probabilidad de observar 4 éxitos, utilizando la fórmula para calcular probabilidades en una binomial, $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$.

Si utilizamos la fórmula de la distribución binomial tendremos:

$P(X = 4) = \binom{7}{4}(1/4)^4(3/4)^3 = 0.0576$

Distribución Binomial en R

En el software R tenemos la siguiente notación para la distribución binomial.

- dbinom(x, size, prob, log = FALSE)
- pbinom(q, size, prob, lower.tail = TRUE, log.p = FALSE)
- qbinom(p, size, prob, lower.tail = TRUE, log.p = FALSE)
- rbinom(n, size, prob)

Distribución Binomial en R

Si X es una variable aleatoria binomial con parámetros $n = 7$ y $p = 1/4$, calcular la probabilidad de observar 4 éxitos utilizando la instrucción adecuada en R

En R podemos utilizar: dbinom(4,7,1/4)

```
## [1] 0.05767822
```

Distribución Binomial en R

¿Qué pasa si calculamos lo siguiente? pbinom(3,7,1/4).

```
## [1] 0.9294434
```

Veamos ahora con:dbinom(0,7,1/4), dbinom(1,7,1/4), dbinom(2,7,1/4), dbinom(3,7,1/4)

```
## [1] 0.1334839
```

```
## [1] 0.3114624
```

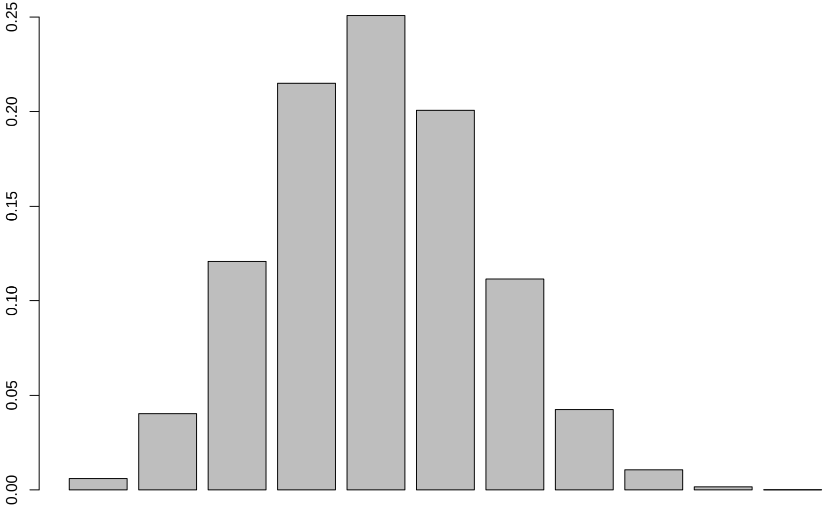
```
## [1] 0.3114624
```

```
## [1] 0.1730347
```

Distribución Binomial en R

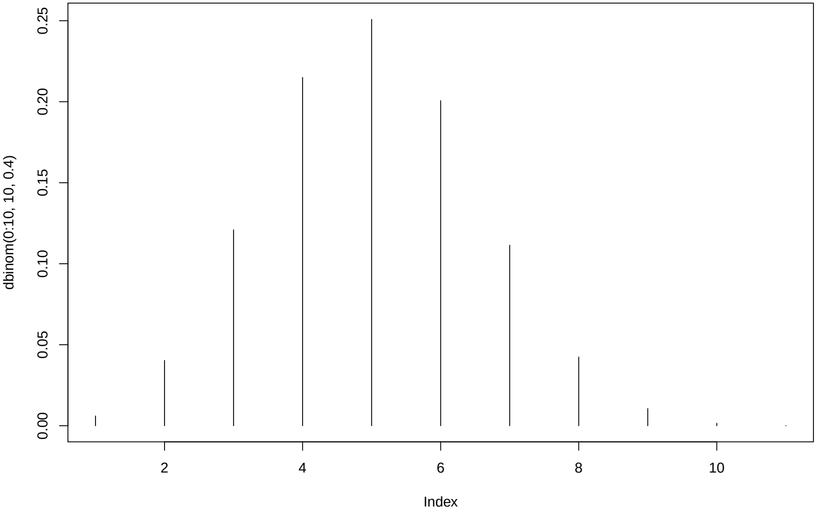
round(dbinom(0:10, size=10, prob=.4),4),

```
##  [1] 0.0060 0.0403 0.1209 0.2150 0.2508 0.2007 0.1115 0.
## [11] 0.0001
```



Otra forma de graficar:

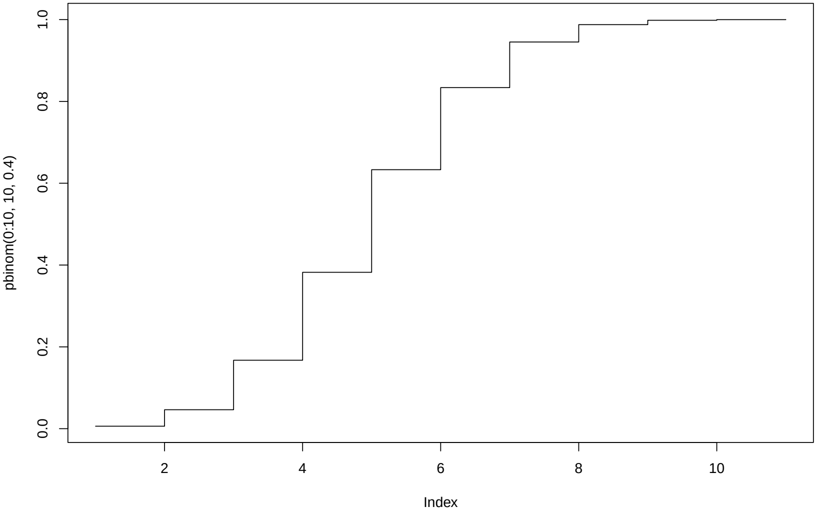
plot(dbinom(0:10,10,0.4),type="h")



Función de distribución acumulada para una binomial

plot(pbinom(0:10,10,0.4),type="s"); 0.006 0.040 0.121 0.215 0.251 0.201 0.111 0.042 0.011 0.002 0.000;

0.006 0.046 0.167 0.382 0.633 0.834 0.945 0.988 0.998 1.000 1.000



Ejercicios

Problema 1. Para una variable aleatoria binomial$X \sim B(16, 0.3)$, calcular las siguientes probabilidades:

- $P(X < 5)$
- $P(X \geq 10)$
- $P(2 < X \leq 12)$
- $P(5 < X \leq 6)$

Distribución Geométrica

Una variable aleatoria que sigue una distribución geométrica permite modelar, realizando repetidos ensayos independientes de Bernoulli, el número de ensayos necesarios hasta obtener el primer éxito, cada ensayo con probabilidad p de éxito.

Decimos que una variable aleatoria X sigue una distribución geométrica, con parámetro p, lo cual podemos denotar como $X \sim Geom(p)$ si su función de probabilidad está dada por

$$P(X = x) = (1-p)^{x-1}p \qquad\qquad x = 1, 2, \ldots, \tag{3}$$

Nota: En ocasiones X se denota como el número de fallas antes de conseguir el primer éxito, y de esa manera, la función de probabilidad se escribe como:

$$P(X = x) = (1-p)^{x}p \qquad\qquad x = 0, 1, 2, \ldots, \tag{4}$$

Distribución Geométrica

Ahora, ¿cómo verificar que la expresión

$$P(X = x) = (1 - p)^{x-1}p = q^{x-1}p \qquad x = 1, 2, \ldots, \tag{5}$$

es una legítima distribución de probabilidad?

Podemos ver que $P(X) \geq 0$ y por otra parte, utilizando el resultado de que una serie geométrica puede ser escrita como $a + ar + ar^2 + ar^3 + ... = \sum_{n=1}^{\infty} ar^{n-1}$ y si $|r| < 1$ la serie converge a $\frac{a}{1-r}$, pues los términos decrecen y en el límite tienden a cero, tenemos que

$\sum_{k=1}^{\infty} P(X = k) = p(1 + q + q^2 + q^3 + \ldots) = p[\frac{1}{1-q}] = 1$

Distribución Geométrica

Si consideramos $p = 0.05$, calculemos:

$$P(X=1) = (1-p)^{x-1}p = (1-p)^0 p = (0.95)^0(0.05) = 0.05$$

$$P(X=2) = (1-p)^1 p = (0.95)(0.05) = 0.0475$$

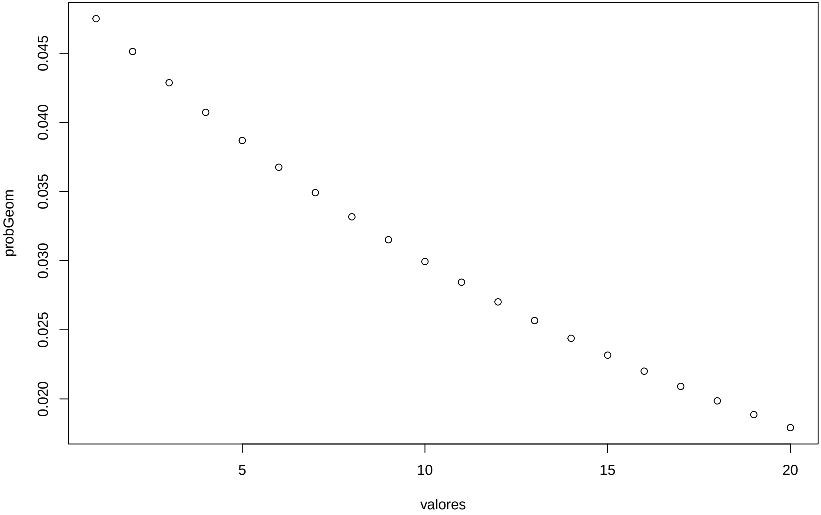
$$P(X=3) = (0.95)^2(0.05) = 0.045125$$

$$P(X=20) = (0.95)^{19}(0.05) = 0.0188$$

```
valores<-seq(1,20,by=1)
probGeom<-dgeom(valores, prob = 0.05)
round(probGeom,4)
```

```
##  [1] 0.0475 0.0451 0.0429 0.0407 0.0387 0.0368 0.0349 0.
## [11] 0.0284 0.0270 0.0257 0.0244 0.0232 0.0220 0.0209 0.
```

Grafiquemos:



Distribución Geométrica en R

Description

Density, distribution function, quantile function and random generation for the geometric distribution with parameter prob.

Usage

- dgeom(x, prob, log = FALSE)
- pgeom(q, prob, lower.tail = TRUE, log.p = FALSE)
- qgeom(p, prob, lower.tail = TRUE, log.p = FALSE)
- rgeom(n, prob)

(tomado de *R*): The geometric distribution with prob=p has density $p(x) = p(1-p)^x$ for $x = 0, 1, 2, ...,\ 0 < p \leq 1$

Distribución Geométrica

Una propiedad de esta variable aleatoria es la “falta de memoria”, que básicamente significa que si X sigue una distribución geométrica, dados dos enteros positivos n y m, se tiene que

$$P(X > n + m | X > n) = P(X > m).$$

Por otra parte, la media y varianza de esta variable aleatoria X son

$$E[X] = \frac{1}{p} \qquad V[X] = \frac{1-p}{p^2}$$

Valor esperado y función de distribución acumulada para $X \sim \text{geom}(p)$

Veamos cómo podemos obtener el valor esperado para X.

$$E(X) = \sum_{k=1}^{\infty} kpq^{k-1} = p\sum_{k=1}^{\infty} \frac{d}{dq}q^k$$

(el intercambiar la diferenciación y la suma es válido pues la serie converge para $|q| < 1$. Ahora,

$$E(X) = p\frac{d}{dq}\sum_{k=1}^{\infty} q^k = p\frac{d}{dq}[\frac{q}{1-q}] = \frac{1}{p}.$$

Algo similar puede hacerse para ver que $V(X) = q/p^2$.

Si $P(X = x) = q^{x-1}p$, con $x = 1, 2, 3, ...$ la función de probabilidad de una variable aleatoria geométrica, entonces, la función de distribución acumulada de X es

$$F(x) = P(X \leq x) = 1 - (1-p)^x = 1 - q^x \qquad x = 1, 2, \ldots$$

Distribución Geométrica: pérdida de memoria

Retomemos la propiedad de "falta de memoria" en la distribución geométrica, que básicamente significa que si X sigue una distribución geométrica, dados dos enteros positivos n y m, se tiene que

$$P(X > n + m|X > n) = P(X > m)$$

Veamos

$$P(X > n) = 1 - (1 - q^n) = (1 - p)^n$$

Luego,

$$P(X > n + m|X > n) = \frac{P(\{X>n+m\} \cap \{X>n\})}{P(X>n)},$$

de esta manera:

$P(X > n + m|X > n) = \frac{P(X>n+m)}{P(X>n)} = \frac{(1-p)^{n+m}}{(1-p)^n}$, para finalmente obtener que:

$$P(X > n + m|X > n) = (1 - p)^m = P(X > m).$$

Distribución Geométrica

En un lote se encuentran N productos nacionales y M productos de origen extranjero. Se selecciona un producto de manera aleatoria, uno a la vez, hasta que se obtenga un producto de origen extranjero. Si suponemos que cada producto seleccionado es reemplazado antes de la siguiente selección, ¿cuál es la probabilidad que sean necesarias exactamente n selecciones?

La distribución geométrica en R

Consideremos que una línea de producción arroja una tasa de 2% de defectuosos. ¿Cuál es la probabilidad de que el primer defectuoso ocurra hasta el sexto artículo seleccionado? X: número de ensayos hasta obtener el primer artículo defectuoso.

$$P(X = 6) = (0.98^5)(0.02) = 0.01807842$$

Para pedirlo en R, escribimos de la siguiente manera, pues la función que considera R es: $P(X = x) = (1 - p)^x p$; esto es, X puede considerarse como el número de fallas antes de que ocurra el primer éxito. Puede solicitarse como $dgeom(x = 5, p = 0.02)$

```
## [1] 0.01807842
```

o bien: $dgeom(5, 0.02)$

```
## [1] 0.01807842
```

La distribución geométrica en R

Por otra parte, la instrucción $pgeom(4, 0.02)$, es la probabilidad de realizar hasta cuatro ensayos antes de observar el primer éxito.

```
## [1] 0.0960792
```

Nótese que sería equivalente a sumar

$$P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5),$$

con la expresión $P(X = x) = (1 - p)^{x-1}p$ que hemos venido utilizando, pero que en el software R conseguiríamos sumando

$$dgeom(0, 0.02) + dgeom(1, 0.02) + \ldots + dgeom(4, 0.02).$$

```
## [1] 0.0960792
```

Ejemplo

Si la probabilidad de que cierto test arroje un resultado positivo es de 0.3,

- ¿Cuál es la probabilidad de que el primer resultado positivo se tenga hasta el sexto ensayo?
- ¿Cuál es la probabilidad de observar menos de 5 reacciones negativas, antes de observar una positiva?

Distribución Hipergeométrica

Consideremos que se tienen N artículos, r de los cuales son defectuosos y $N-r$ no defectuosos. Seleccionamos aleatoriamente n de ellos, donde $0 \leq k \leq r$ y $0 \leq n-k \leq N-r$. Denotemos por X el número de artículos defectuosos que resultan en la muestra. Si

$$P(X=k) = \frac{\binom{r}{k}\binom{N-r}{n-k}}{\binom{N}{n}} \tag{6}$$

una variable aleatoria discreta que tiene la función de probabilidad dada en (6), se dice que tiene una distribución hipergeométrica, que es utilizada para muestrear, sin reemplazamiento, de una población finita.

Distribución hipergeométrica

El valor esperado para una variable aleatoria X que sigue una distribución hipergeométrica es:

$$\mu_X = E(X) = \frac{nr}{N},$$

mientras que la varianza está dada por:

$$\sigma_X^2 = V(X) = n(\frac{r}{N})(\frac{N-r}{N})(\frac{N-n}{N-1}),$$

y recordemos que la desviación estándar $\sigma_X = \sqrt{V(X)}$.

Nota: Cuando r y N son relativamente grandes con respecto a n, la función de probabilidad de de X puede ser aproximada por una variable aleatoria binomial con parámetros n y p, donde $p = m/N$.

Ejemplo

Cierto tipo de motores se embarcan en lotes de 60. Antes de que un lote sea aceptado, un ingeniero selecciona 6 de estos motores y los inspecciona. Si ninguno de ellos es defectuosos, el lote es aceptado. Si uno o más son defectuosos, el lote entero se inspecciona.

Supongamos que en realidad hay tres motores defectuosos en el lote. ¿Cuál es la probabilidad de que tenga que inspeccionarse el lote entero? Defina X.

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \frac{\binom{3}{0}\binom{57}{6}}{\binom{60}{6}} = 1 - 0.7248393 = 0.2752$$

Distribución hipergeométrica en R

Para calcular probabilidades en una distribución hipergeométrica, con el software R, podemos utilizar la instrucción $dhyper(x, m, n, k)$ y para probabilidades acumuladas hasta un cierto valor q, utilizamos $phyper(q, m, n, k, lower.tail = TRUE)$, donde m número de artículos defectuosos, n son los no-defectuosos, k el número de items que se extraen de la población. (En nuestra notación tendríamos $dhyper(k, r, N - r, n)$

Veamos el ejercicio anterior y calculemos: $1 - dhyper(0, 3, 57, 6)$:

```
## [1] 0.2751607
```

¿Cómo utilizar la probabilidad acumulada? Calculemos $phyper(0, 3, 57, 6, lower.tail = FALSE)$

```
## [1] 0.2751607
```

Distribución hipergeométrica: Ejemplo

Siete computadoras son elegidas, aleatoriamente, de entre un embarque de 240 computadoras, de las cuales 15 son defectuosas. Encontrar la probabilidad de que 4 de las computadoras seleccionadas resulten defectuosas. Definir X.

Solución: En nuestra notación tendríamos $dhyper(k, r, N - r, n)$, esto es: dhyper(4,15,225,7)

```
## [1] 0.0003069143
```

-La probabilidad de que al menos 1 de las computadoras sea defectuosa. Recordemos que $phyper(k, r, N - r, n, lower.tail = TRUE)$, nos proporciona la probabilidad acumulada hasta un valor k. Necesitamos calcular $P(X \geq 1) = 1 - P(X = 0)$, mediante: $1 - dhyper(0, 15, 225, 7)$,

```
## [1] 0.3672717
```

Tarea: Ejemplo el uso de la distribución hipergeométrica en estudios de captura-recaptura.

Distribución Poisson

Decimos que una variable aleatoria X tiene una distribución de Poisson, con parámetro de escala λ, positivo, lo cual denotamos como $X \sim \text{Poisson}(\lambda)$, si su función de probabilidad está dada por

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!} \qquad\qquad k = 0, 1, 2, \ldots, n, \ldots$$

La distribución de Poisson puede ser utilizada para modelar el número de eventos en un cierto intervalo, asociado con un proceso que involucra aleatoriedad en el espacio, o en el tiempo.

Distribución Poisson

Observemos que:

$$\sum_{k=0}^{\infty} P(X = k) = \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}e^{\lambda} = 1,$$

pues recordemos que

$$\sum_{n=0}^{\infty} \frac{x^n}{n!} = e^x$$

Ejemplos

Algunos ejemplos de variables aleatorias que pueden modelarse con una Poisson, son:

- Número de errores tipográficos en una página, grupo de páginas o en un libro.
- Número de clientes que llegan a una fila, en una hora.
- Número de llamadas telefónicas en un cierto período de tiempo.
- Número de partículas emitidas por un material radioactivo.

Nota: la distribución de Poisson puede utilizarse para aproximarse a una distribución binomial, cuando n es grande y p es pequeña.

Valor esperado y varianza

Teorema. Si X sigue una distribución Poisson, con parámetro λ, entonces:

$$E(X) = \lambda, \qquad y \qquad V(X) = \lambda$$

Esto es,

$$E(X) = \sum_{k=0}^{\infty} kP(X = k) = \sum_{k=0}^{\infty} k\frac{e^{-\lambda}\lambda^k}{k!} = \sum_{k=1}^{\infty} \frac{e^{-\lambda}\lambda^k}{(k-1)!},$$

luego, haciendo el cambio de variable $s = k - 1$ tenemos que:

$$E(X) = \lambda \sum_{s=0}^{\infty} \frac{e^{-\lambda}\lambda^s}{(s)!} = \lambda.$$

Valor esperado y varianza

Veamos ahora el cálculo de la varianza:
$V(X) = E(X^2) - [E(X)]^2$, y sabemos que $E(X) = \lambda$, entonces, sólo debemos calcular:

$$E(X^2) = \sum_{k=0}^{\infty} k^2 \frac{e^{-\lambda}\lambda^k}{k!} = \sum_{k=1}^{\infty} k \frac{e^{-\lambda}\lambda^k}{(k-1)!},$$

nuevamente hacemos $s = k - 1$, esto es, $k = s + 1$,

$$E(X^2) = \sum_{s=0}^{\infty} (s+1) \frac{e^{-\lambda}\lambda^{s+1}}{(s)!} = \lambda \sum_{s=0}^{\infty} s \frac{e^{-\lambda}\lambda^s}{s!} + \lambda \sum_{s=0}^{\infty} \frac{e^{-\lambda}\lambda^s}{s!}$$

$$E(X^2) = \lambda^2 + \lambda,$$

de esta manera:

$$V(X) = E(X^2) - [E(X)]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

Ejemplos

Suponga que el número promedio de carros abandonados, por semana, en cierta carretera es de 2.2. Calcule la probabilidad de que:

- No abandonen ningún carro la próxima semana.

  Si consideramos X como el número de carros abandonados por semana, y $\lambda = 2.2$ entonces, como:

  $$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!},$$

  tendremos que

  $$P(X = 0) = \frac{e^{-2.2}2.2^0}{0!} = 0.1108032,$$

Ejemplos

- Al menos haya dos carros abandonados la próxima semana.

$$P(X \geq 2) = 1 - P(X \leq 1) = 1 - P(X = 0) - P(X = 1)$$

$$P(X \geq 2) = 1 - \frac{e^{-2.2} 2.2^0}{0!} - \frac{e^{-2.2} 2.2^1}{1!}$$

$P(X \geq 2) = 1 - 0.1108032 - 0.2437669 = 0.6454299.$

¿Cómo podemos hacer lo anterior en el software R?

- No abandonen ningún carro la próxima semana. Usamos la instrucción: $dpois(0, 2.2)$

```
## [1] 0.1108032
```

- Al menos haya dos carros abandonados la próxima semana, con la instrucción: $1 - ppois(1, 2.2)$

```
## [1] 0.6454299
```

Ejemplos

El número de partículas radioactivas emitidas cada minuto, por un meteorito, se registra con un contador Geiger y se ha encontrado que el número promedio por minuto es de 3.5 partículas. Utilizando una distribución de Poisson, encuentre la probabilidad de que, en cualquier minuto:

- No se emitan partículas radioactivas.
- Se emitan dos partículas radioactivas.
- Se emitan al menos 5 partículas radioactivas.

Aproximación Poisson a una distribución binomial

Teorema. Sea X una variable aleatoria con distribución binomial, con parámetro p (basado en n repeticiones del experimento), esto es:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

y suponga que cuando $n \to \infty$, $np = \lambda$ (constante), o equivalentemente $n \to \infty, \quad p \to 0$, de manera que $np \to \lambda$. Bajo estas condiciones se tiene:

$$\lim_{n\to\infty} P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!},$$

la distribución Poisson con parámetro λ.

Ejemplos

El 5% de cierto tipo de focos manufacturados por una compañía, resultan defectuosos. El ingeniero encargado del control de calidad diariamente selecciona una muestra aleatoria de 100 focos. Denotemos por X el número de focos defectuosos en la muestra.

- Cuál es la probabilidad de que la muestra contenga a lo más tres focos defectuosos? Si realizamos el cálculo con la binomial, considerando se cumplen los supuestos de ésta, tendremos que
  $P(X \leq 3) = \sum_{k=0}^{3} \binom{100}{k} (0.05)^k (.95)^{100-k} = 0.2578387$

Si consideramos que $np \to \lambda$, tendremos que $\lambda = 100 * 0.05 = 5$. Luego, con la aproximación Poisson a la binomial tendremos que:

$P(X \leq 3) = \sum_{k=0}^{3} \frac{e^{-5} 5^k}{k!} = 0.2650259$

¿Cómo solicitaría lo anterior en R?

Distribución Binomial Negativa: Antecedentes

Recordemos que la distribución binomial nos permite calcular la probabilidad de observar k exitos en n ensayos independientes de Bernoulli, donde la probabilidad de ensayo permanece constante y fija en p.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{7}$$

Distribución Binomial Negativa: Antecedentes

¿Cómo obtuvimos la fórmula anterior? Consideremos una secuencia en particular, de n ensayos de Bernoulli, donde se tengan k éxitos, por ejemplo, si denotemos por F al fracaso y E éxito, entonces la secuencia:$E, E, F, E, F, \ldots, E$, a la que asociaremos la probabilidad $p \cdot p \cdot (1-p) \cdot p \cdot (1-p) \ldots p$. Si reagrupamos, podemos escribir la probabilidad anterior como: $p^k(1-p)^{n-k}$ Esto es, la probabilidad de que una secuencia en particular tenga k éxitos. Cuando se trabajan todas las posibles secuencias, entonces tenemos:
$\binom{n}{k}$, y la fórmula en (7) queda clara.

Distribución Binomial Negativa: Antecedentes

Consideremos la probabilidad de r éxitos y x fallas en un orden específico. Por ejemplo, la secuencia

$$FFEFFEFFE,$$

tenemos entonces que $r = 3$ y $x = 6$, y la probabilidad estaría dada por:

$$p^r(1-p)^x$$

y tenemos $\binom{x+r-1}{r-1}$ diferentes secuencias asociadas con x fallas antes del rth éxito.

Distribución binomial negativa

La distribución Binomial negativa, también conocida como distribución de Pascal, puede ser utilizada para modelar el número de fallas x antes del rth éxito, en ensayos independientes y repetidos de Bernoulli, cada uno con probabilidad p.

La notación $X \sim NB(r, p)$ se utiliza para denotar una variable aleatoria X que tiene una distribución binomial negativa, con r un entero positivo y $0 < p < 1$ un parámetro real, cuya función de probabilidad está dada por:

$$P(X = x) = \binom{x + r - 1}{r - 1} p^r (1 - p)^x \qquad x = 0, 1, 2, \ldots.$$

Distribución binomial negativa

Las suposiciones para una distribución binomial negativa son:

El experimento consiste de una sucesión de ensayos independientes.

Cada ensayo resulta en éxito o fracaso.

La probabilidad de éxito es constante de ensayo a ensayo, para $i = 1, 2, 3, \ldots$

El experimento continúa hasta obtener un total de r éxitos, donde r es un entero positivo.

Distribución binomial negativa

Es posible que también encontremos que una variable aleatoria X que denota el número de ensayos hasta de alcanzar r éxitos se defina como una binomial negativa. En ese caso, si se efectuaron x ensayos hasta alcanzar r éxitos, se tendrían $x - r$ fracasos.

De cuántas maneras podemos obtener $r - 1$ éxitos en $x - 1$ ensayos?

$$P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r},$$

con $x \geq r$

La media y varianza de X están dadas por: $E[X] = r/p$ y $V(X) = rq/p^2$

Media y Varianza: binomial, Poisson, binomial negativa

Si X se distribuye binomialmente y representa número de éxitos en n ensayos independientes de Bernoulli, donde p es la probabilidad de éxito y permanece constante de ensayo a ensayo, hemos visto que

$E(X) = np$ y $V(X) = np(1-p)$

Por otro lado, cuando la variable aleatoria X se distribuye Poisson con parámetro λ, representando el número de veces que un evento ocurre en un intervalo de tiempo (espacio), se tiene que $E(X) = \lambda$ y $V(X) = \lambda$.

En la lámina anterior se tiene que si X es una v.a. que se distribuye binomial negativa, con p la probabilidad de éxitos y modelando el número de fallas hasta alcanzar r éxitos:

$E[X] = r/p$ y $V[X] = rq/p^2$

Binomial negativa en R

```
dnbinom(x, size, prob, mu, log = FALSE)
pnbinom(q, size, prob, mu, lower.tail = TRUE, log.p = FALSE)
qnbinom(p, size, prob, mu, lower.tail = TRUE, log.p = FALSE)
rnbinom(n, size, prob, mu)
```

Arguments

`x`	vector of (non-negative integer) quantiles.
`q`	vector of quantiles.
`p`	vector of probabilities.
`n`	number of observations. If `length(n) > 1`, the length is taken to be the number required.
`size`	target for number of successful trials, or dispersion parameter (the shape parameter of the gamma mixing distribution). Must be strictly positive, need not be integer.
`prob`	probability of success in each trial. `0 < prob <= 1`.
`mu`	alternative parametrization via mean: see 'Details'.
`log, log.p`	logical; if TRUE, probabilities p are given as log(p).
`lower.tail`	logical; if TRUE (default), probabilities are $P[X \leq x]$, otherwise,

Ejemplo

Supongamos que la probabilidad de que usted venda la licencia de un software es de 0.3. Cierto día inicia la visita de clientes y decide detener estas visitas cuando tenga 5 licencias vendidas. ¿Cuál es la probabilidad de que termine su venta en el octavo cliente?

Consideremos que X es el número de visitas hasta vender 5 licencias, y como $p = 0.3$, utilizando:

$$P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r},$$

tendremos que:

$$P(X = 8) = \binom{7}{4} 0.3^5 (.7)^3$$

En R podríamos calcular lo anterior manualmente por medio del siguiente cálculo $\textit{choose}(7, 4) * (0.3^5) * (0.7)^3$, con lo cual obtenemos 0.02917215.

Distribución binomial negativa en R

Para el problema anterior tenemos que: X es el número de ensayos antes de alcanzar 5 éxitos.

En R, la instrucción $dnbinom(x, size, prob)$, donde:

- $size$ es el numero de ensayos exitosos.
- x es el número de fallas que ocurren, en una sucesión de ensayos de Bernoulli, antes de alcanzar un cierto número de éxitos.
- p es la probabilidad de éxito en cada ensayo.

Considerando lo anterior tendríamos que la instrucción es $dnbinom(3, 5, 0.3)$

```
## [1] 0.02917215
```

Generación de números aleatorios en una binomial negativa

En *rnbinom(n, size, prob)*, n es el tamaño de muestra (cuántos números aleatorios), size es el número de ensayos exitosos (nuestra r) y p es la probabilidad de éxito.

```
r=4
p=0.2
rnbinom(10,r,p)
```

```
##  [1] 13 34 12 22 35  4  7 22  4 15
```

```
media<-r/p
media
```

```
## [1] 20
```

```
varianza<-r*(1-p)/p^2
varianza
```

```
## [1] 80
```

Cálculo de media y varianza a través de simulaciones

Qué pasa si por medio de simulaciones tratamos de obtener estimaciones de la media y varianza, cuando $r = 4$ y $p = 0.2$?

```
datos<-rnbinom(10000,r,p)
head(datos)
```

```
## [1]  3  7 19 27 10 14
```

```
mean(datos)
```

```
## [1] 15.9933
```

```
var(datos)
```

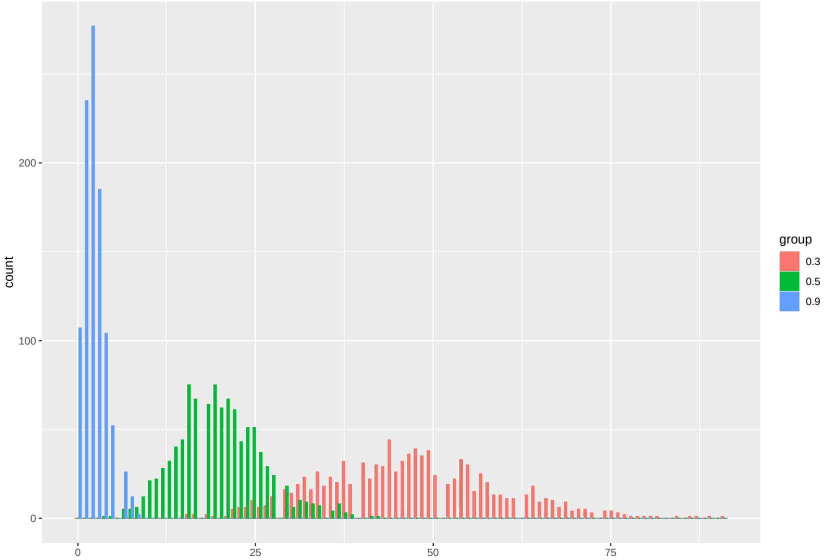
```
## [1] 80.57251
```

Simulación de Mil datos con $r = 20$, diferente p

```
rand1<-rnbinom(1000,20,p=0.3)
rand2<-rnbinom(1000,20,p=0.5)
rand3<-rnbinom(1000,20,p=0.9)
x<-c(rand1,rand2,rand3)
group<-c(rep("0.3",1000),rep("0.5",1000),rep("0.9",1000))
dfr<-data.frame(x,group=group)
```

Comparación distribuciones, $r = 20$, diferente p

```
ggplot(dfr,aes(x,fill=group,colour=group))+
  geom_histogram(position = "dodge",bins=100)
```



Simulación de mil datos con diferente r, $p = 0.2$

```
prob11<-rnbinom(1000,10,p=0.2)
prob21<-rnbinom(1000,20,p=0.2)
prob31<-rnbinom(1000,50,p=0.2)
x1<-c(prob11,prob21,prob31)
group1<-c(rep("r=10",1000),rep("r=20",1000),rep("r=50",1000
dfr<-data.frame(x1,group=group1)
```

Comparación distribuciones con diferente r, $p = 0.2$

```
ggplot(dfr,aes(x1,fill=group1,colour=group1))+
  geom_histogram(position = "dodge", bins = 100)
```