Variables aleatorias

Curso 2022-23

13 de Febrero al 17 de Febrero de 2022

Variables aleatorias

Variables aleatorias

- El concepto de variable aleatoria es un modo de unir técnicas que nos han aparecido en probabilidad y estadística descriptiva
- Entre otros conceptos:
  - De probabilidad
    - Sucesos y probabilidades asociados a ellos
    - Probabilidad, independencia, probabilidad condicionada y total, regla de Bayes
  - De estadística descriptiva
    - Medidas centrales y de dispersión como la media y la desviación típica e histogramas
    - Percentiles y datos atípicos
- Para poder usar todas estas técnicas, los sucesos a estudiar deben de ser numéricos

Variables aleatorias

- Partimos de un experimento aleatorio cualquiera
- Los sucesos no tienen por qué ser números
- Nos fijamos en una característica numérica de estos sucesos
- La variable aleatoria permite estudiar esa característica numérica
- Más formalmente
  - Una variable aleatoria es asignar a cada resultado del experimento un número
  - En vez de estudiar probabilidades de los sucesos del experimento estudiamos probabilidades de los números asociados
  - Matemáticamente, una variable aleatoria es una función $X : \Omega \to R$

Variables aleatorias discretas

Variables aleatorias discretas

- Una variable aleatoria discreta viene dada por
  - Una lista de los valores que puede tomar esa variable
  - Una lista de probabilidades de esos valores ya calculadas (con cualquier técnica)
- A partir de aquí
  - Al dibujar los valores junto a sus probabilidades se obtiene el equivalente al barplot o histograma de densidades
  - Faltaría
    - Definir media y varianza en función de las dos listas
    - Percentiles

Media y varianza de una variable aleatoria discreta

- Partimos de una variable aleatoria discreta X dada por:
  - Una lista de posibles valores $(x_1, \cdots, x_n)$
  - Una lista de probabilidades de esos valores ya calculadas $p_i = P(X = x_i)$
  - Suele denotarse $f(x_i) = P(X = x_i)$
- Media o esperanza de X

$$m = E[X] = x_1 \cdot p_1 + \cdots + x_n \cdot p_n$$

- En general, si $g(X)$ es cualquier función de la variable X

$$E[g(X)] = g(x_1) \cdot p_1 + \cdots + g(x_n) \cdot p_n$$

- Varianza

$$E[(X - m)^2] = (x_1 - m)^2 \cdot p_1 + \cdots + (x_n - m)^2 \cdot p_n$$

Función de distribución y percentiles

- Se llama función de distribución de X ,y se denota por F(x), a la función que asigna a cada valor x su percentil
- Una forma diferente de verlo es que F(x) asigna a x la probabilidad de ser menor o igual que ese valor

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

- Una forma sencilla de calcular F(x)
  - Calcular una lista con probabilidades acumuladas usando la función cumsum() en R
  - Para cualquier valor x_i de la lista de valores de X, el valor de $F(x_i)$ es el valor correspondiente en la lista calculada anteriormente
  - Para cualquier otro valor
    - Buscar en qué intervalo está de entre

$$(-\infty, x_1), [x_1, x_2), \cdots, [x_{n-1}, x_n), [x_n, \infty)$$

    - Si es menor que x_1 entonces $F(x) = 0$
    - En los demás casos, si está en el intervalo $[a, b)$, entonces $F(x) = F(a)$

Variables aleatorias continuas

Variables aleatorias continuas

- Una variable aleatoria se dice continua si puede tomar todos los valores de un intervalo.
- Una variable aleatoria continua viene dada por:
  - Un intervalo $[a, b]$ de posibles valores (puede ser abierto, etc)
  - Una función $f(x)$ (llamada de densidad) que nos indica qué regiones del intervalo son más probables que otras
- Esta función debe cumplir
  - $f(x) \geq 0, \forall x \in [a, b]$
  - $\int_a^b f(x)dx = 1$
- A partir de aquí
  - Al dibujar la función $f(x)$ se obtiene el equivalente al histograma de densidades

Media, varianza y esperanza para variables aleatorias continuas

- Son parecidas a las variables aleatorias discretas. Ahora:
  - Los valores vienen representados por la variable x
  - Las probabilidades por la función de densidad $f(x)$
  - Las sumas se convierten en integrales
- Con estos cambios
  - Media o esperanza de X: $m = E[X] = \int_a^b x \cdot f(x)dx$
  - Esperanza de una función $g(X)$: $E[g(X)] = \int_a^b g(x) \cdot f(x)dx$
  - Varianza: $v = E[(X - m)^2] = \int_a^b (x - m)^2 \cdot f(x)dx$

Función de distribución de una variable aleatoria continua

- Tiene el mismo papel que en variables aleatorias discretas
  - $F(x) = P(X \leq x)$
  - Por tanto $F(x)$ vale
    - 0 xi $x < a$
    - 1 si $x > b$
    - $\int_a^x f(t)dt$ si $a \leq x \leq b$
- Se utiliza para
  - Calcular probabilidades en subintervalos
  - Calcular percentiles

Calcular probabilidades para una variable aleatoria

Cálculo de probabilidades si la variable es discreta

- En una variable aleatoria discreta
  - La probabilidad de estar en un intervalo es sumar las probabilidades de los valores del intervalo
  - Ejemplo
    - Una variable aleatoria toma valores $\{1, 2, \cdots, 10\}$ con probabilidades $P(X = i) = \frac{i^2}{385}$
    - En R tendríamos las listas $val = 1:10$, $prob = val^2/385$
    - Para calcular $P(3 < X <= 7)$ sumamos las probabilidades de 4, 5, 6, 7 que son los valores que lo cumplen
    - En R: $sum(prob[3 < val \ \& \ val <= 7])$

Cálculo de probabilidades si la variable es continua

- En una variable aleatoria continua
  - La probabilidad de estar en un intervalo es sumar las probabilidades de los valores del intervalo
  - Ejemplo
    - Una variable aleatoria toma valores en el intervalo $[0, 3]$ con función de densidad

$$f(x) = \begin{cases} 0, & \text{si } x < 0 \\ \frac{x^2}{9}, & \text{si } 1 \leq x \leq 3 \\ 0, & \text{si } x > 3 \end{cases}$$

    - $P(1 < X < 2) = \int_1^2 f(x)dx$
- Por tanto, en una variable aleatoria continua, $P(X = a) = \int_a^a f(x)dx = 0$
- No es cierto que $P(X = a) = f(a)$

Propiedades importantes

Variables independientes

- Sean X e Y dos variables aleatorias discretas. Se dice que X e Y son independientes si los sucesos $P(X = x)$ y $P(Y = y)$ son independientes para cualquier posible valor x de X y cualquier posible valor y de Y
- Si X e Y son dos variables aleatorias discretas, se dice que X e Y son independientes si los sucesos $P(X \leq x)$ y $P(Y \leq y)$ son independientes para cualquier posible par de números reales x e y.

Propiedades de la media y la varianza

- Sen X e Y dos variables aleatorias, a, b dos números
- Propiedades de la media
  - $E[a \cdot X + b] = a \cdot E[X] + b$
  - $E[X + Y] = E[X] + E[Y]$
  - Hay que tener cuidado con otras operaciones, por ejemplo $E[X \cdot Y] \neq E[X] \cdot E[Y]$
- Propiedades de la varianza
  - $Var(a \cdot X + Y) = a^2 \cdot Var(X)$
  - Si X e Y son independientes, $Var(X + Y) = Var(X) + Var(Y)$

Modelos de probabilidad

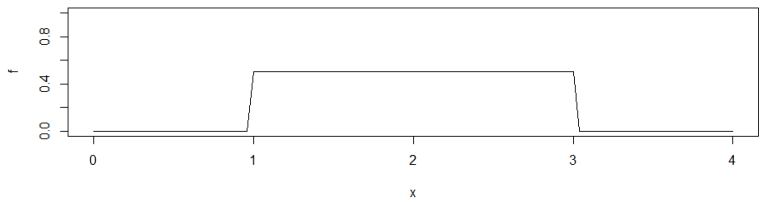
Modelos uniformes

Modelo uniforme discreto

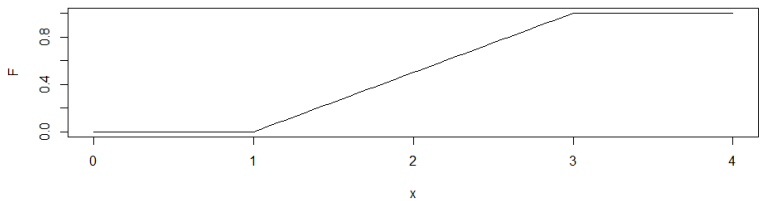
- Tenemos un conjunto finito de posibles resultados $\{x_1, \cdots, x_n\}$
- Suponemos que todos tienen la misma probabilidad $P(X = x_i) = \frac{1}{n}$
  - La función de probabilidad f es constante
  - La función de distribución viene dada por $F(x_i) = \frac{i}{n}$
  - La media es la media de los valores
  - La varianza es la varianza de los valores
  - Los percentiles son los percentiles de los valores

Modelo uniforme continuo

- Posibles valores en un intervalo $[a, b]$
- Todos tienen 'la misma probabilidad' luego f es constante
- Debe ser $f(x) = \frac{1}{b-a}$, $\forall x \in [a, b]$



- La función de distribución viene dada por $F(x) = \frac{x-a}{b-a}$



Modelo uniforme continuo

- Media, varianza y percentiles
  - La media es $\frac{a+b}{2}$
  - La varianza es $\frac{(b-a)^2}{12}$
  - El percentil $p\%$ es $a + p \cdot (b - a)$
- En la mayor parte de los lenguajes de programación se corresponde a la función random()
- En R se simula con runif(, a , b)

Suceso de Bernouilli: modelos binomial y geométrico

Suceso de Bernouilli

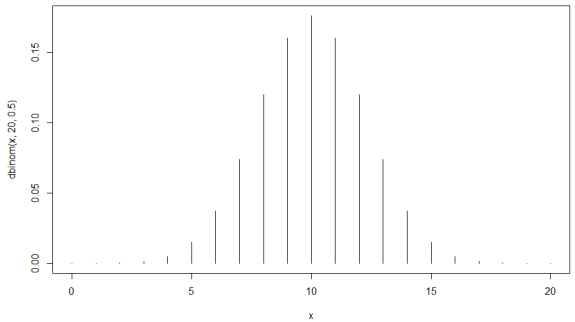
- Tenemos un experimento del que solamente nos interesa un posible resultado
- Conocemos su probabilidad p
- Lo modelamos con una variable aleatoria con dos valores
  - 0 si no se da ese suceso
  - 1 si se da el suceso
- Por tanto $P(X = 1) = p$ y $P(X = 0) = 1 - p$
- Su media es p y su varianza es $p \cdot (1 - p)$
- No tiene sentido calcular función de distribución o percentiles

Modelo binomial

- Tenemos un suceso de Bernouilli y lo repetimos un determinado número de veces n
- Suponemos que esas repeticiones del experimento son independientes (es decir, que la probabilidad p permanece constante)
- La variable aleatoria $X =$ 'número de veces que ocurre el suceso en esos n intentos' se llama variable binomial con parámetros n y p
- Se denota por $B(n, p)$
- En R se escribe $binom(-, n, p)$

Modelo binomial

- Valores entre 0 y n con $P(x = i) = \binom{n}{i} p^i (1-p)^{n-i}$
- En R se calcula como $dbinom(i, n, p)$



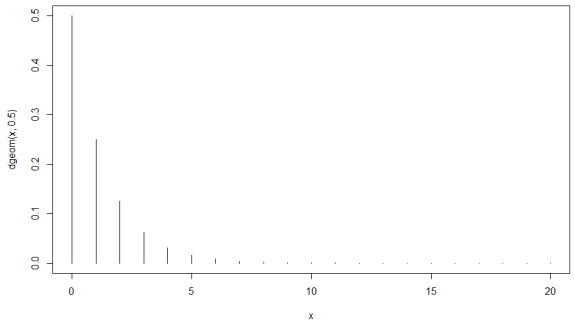
- Las órdenes $pbinom(-, n, p)$ y $qbinom(-, n, p)$ permiten calcular probabilidades acumuladas y percentiles
- Su media es $n \cdot p$ y su varianza es $n \cdot p \cdot (1-p)$
- Se simula con $rbinom(-, n, p)$

Modelo geométrico

- Tenemos un suceso de Bernouilli y lo repetimos hasta que ocurre el suceso que nos interesa
- Suponemos que esas repeticiones del experimento son independientes (es decir, que la probabilidad p permanece constante)
- La variable aleatoria $X =$ 'número de repeticiones antes de que ocurra el suceso' se llama variable geométrica con parámetro p
- Se denota por $G(p)$
- En R se escribe $geom(-, p)$

Modelo geométrico

- Valores $0, 1, 2 \cdots$ con $P(x = i) = p(1-p)^i$
- En R se calcula como $dgeom(i, p)$

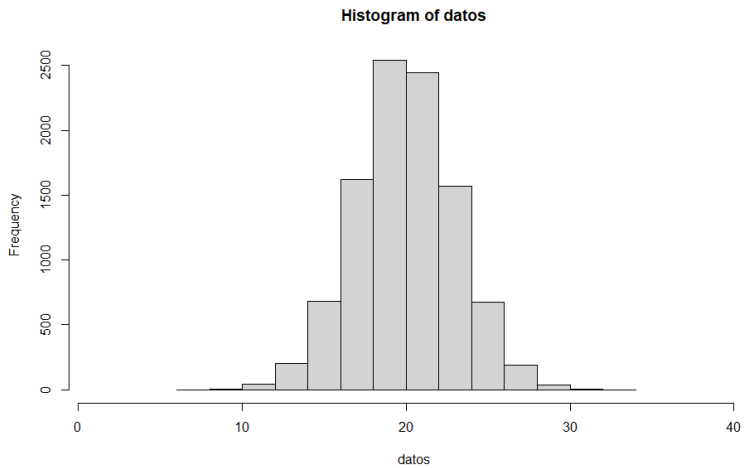


- Las órdenes $pgeom(-, p)$ y $qgeom(-, p)$ permiten calcular probabilidades acumuladas y percentiles
- Su media es $\frac{1-p}{p}$ y su varianza es $\frac{1-p}{p^2}$
- Se simula con $rgeom(-, p)$

Modelo normal

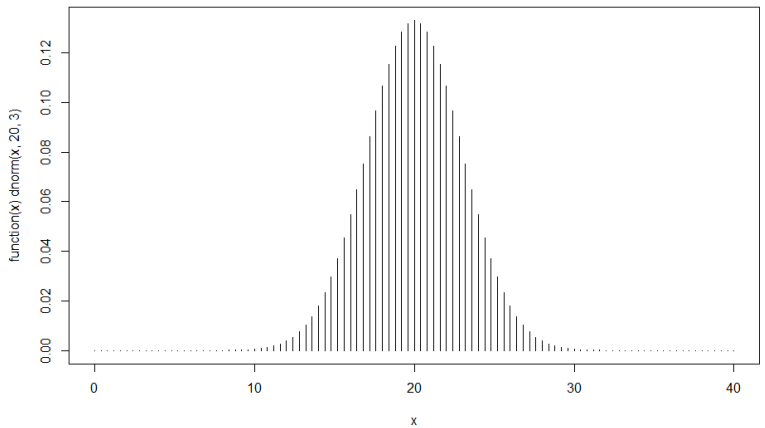
Modelo normal

- Cuando medimos una cantidad solemos cometer pequeños errores
  - Las medidas obtenidas suelen ser parecidas a la medida correcta
  - Habitualmente es igual de fácil equivocarse por exceso o por defecto
  - Es menos probable equivocarse mucho que poco
- La variable normal se utiliza para modelar estas situaciones



Variable normal

- Una variable X se dice que es una variable normal con parámetros m y s si
  - Puede tomar valores entre $-\infty$ y $+\infty$
  - Su función de densidad es $f(x) = \frac{1}{s\sqrt{2\pi}} e^{-\frac{x-m}{2s^2}}$



Variable normal

- Se escribe $X \simeq N(m, s)$
- Su media es m y su desviación típica es s
- En R se usan las órdenes $dnorm(-, m, s)$, $pnorm(-, m, s)$, $qnorm(-, m, s)$ y $rnorm(-, m, s)$ para calcular las funciones de densidad, de distribución, calcular percentiles o simular la variable

Propiedades importantes de las variables normales

- Si $X \simeq N(m, s)$ y $r \in \mathbb{R}$ es un número entonces

$$r \cdot X \simeq N(r \cdot m, r \cdot s)$$

- Si $\{X_1, \cdots, X_n\}$ son variables independientes con cada $X_i \simeq N(m_i, s_i)$, entonces

$$X_1 + \cdots + X_n \simeq N(m_1 + \cdots + m_n, \sqrt{s_1^2 + \cdots + s_n^2})$$

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n} \simeq N(\frac{m_1 + \cdots + m_n}{n}, \frac{\sqrt{s_1^2 + \cdots + s_n^2}}{n})$$

Variables asociadas a la variable normal

- Vamos a definir dos variables asociadas a una variable normal
- Variable chi cuadrado
  - Sean $X_1, \cdots, X_n$ variables independientes y cada $X_i \simeq N(0,1)$. La variable $\chi^2_n = X_1^2 + \cdots X_n^2$ se llama variable chi-cuadrado con parámetro n
  - En R se escribe $chisq(, n)$
- Variable t de Student
  - Se llama variable t de Student con parámetro n a $t_n = \frac{N(0,1)}{\sqrt{(\frac{\chi^2_n}{n})}}$
  - En R se escribe $t(, n)$

Proceso de Poisson: modelo de Poisson y exponencial

Proceso de Poisson

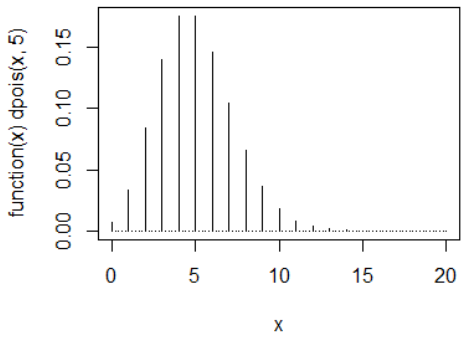
- Estamos interesados en un suceso del que no se conoce la probabilidad. Queremos estudiar qué ocurre en un intervalo de tiempo (o espacio)
- El proceso se dice de Poisson si
  - A la vez solamente ocurre un evento (los tiempos de ocurrencia de los eventos pueden, aún así, estar muy cerca unos de otros).
  - El número de veces que ocurre el suceso en un intervalo es independiente del número de veces que ocurre en cualquier otro intervalo.
  - El número de veces que ocurre el suceso en un intervalo crece de forma aproximadamente proporcional con el tiempo.
- Este proceso viene determinado por cualquiera de los dos datos siguientes
  - El número de veces que ocurre este suceso, en media, en cualquier intervalo finito (llamado λ)
  - El tiempo medio que tarda en ocurrir ese suceso

Modelo de Poisson

- Tenemos un proceso de Poisson con parámetro λ en un intervalo
- A la variable X = 'Número de veces que ocurre el suceso en ese intervalo' se le llama variable de Poisson
- Se denota por $Pois(\lambda)$
- En R se escribe $pois(-, p)$

Modelo de Poisson

- Valores $0, 1, 2 \cdots$ con $P(x = i) = e^{-\lambda} \frac{\lambda^i}{i!}$
- En R se calcula como $dpois(i, lambda)$



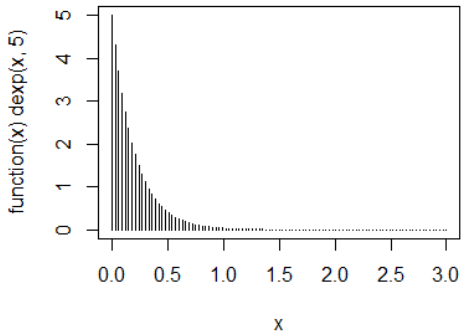
- Las órdenes $ppois(-, lambda)$ y $qpois(-, lambda)$ permiten calcular probabilidades acumuladas y percentiles
- Su media es λ y su varianza es λ
- Se simula con $rpois(-, lambda)$

Modelo exponencial

- Tenemos un proceso de Poisson con parámetro λ en un intervalo
- A la variable X = 'Tiempo transcurrido hasta que vuelva a ocurrir el suceso' se le llama variable exponencial
- Se denota por $E(\lambda)$
- En R se escribe $exp(-, p)$

Modelo exponencial

- Valores $[0, +\infty)$ con $f(x) = \lambda e^{-\lambda x}$
- En R se calcula como $dexp(-, lambda)$



- Las órdenes $pexp(-, lambda)$ y $qexp(-, lambda)$ permiten calcular probabilidades acumuladas y percentiles
- Su media es $\frac{1}{\lambda}$ y su varianza es $\frac{1}{\lambda^2}$
- Se simula con $rexp(-, lambda)$