Variables Aleatorias III. Distribuciones notables.

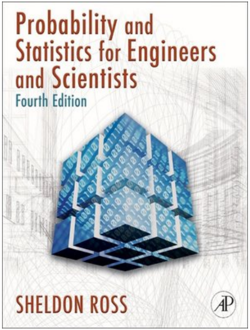
Estadística, Grado en Sistemas de Información

Constantino Antonio García Martínez

Universidad San Pablo Ceu

Contenidos

References



S. Ross. Introduction to Probability and Statistics for Engineers and Scientists. Chapter 5.

C.D. Barr, D.M. Diez, M. Çetinkaya-Rundel. OpenIntro Statistics. Chapters 3.

Distribuciones discretas

Distribuciones discretas

Ensayos de Bernouilli

Jacob Bernouilli.

Distribución binomial
En un experimento que repetimos varias veces, llamamos a cada repetición **ensayo**. Si los ensayos son independientes entre sí y solo hay dos resultados posibles (**éxito** y **fracaso**), llamamos a estos experimentos **ensayos de Bernouilli**. Existen muchas VAs notables que surgen del estudio de experimentos de Bernouilli:

- Distribución Bernouilli y binomial.
- Distribución geométrica.
- Distribución binomial negativa.

Ejercicio: Distribución binomial
Se tira una moneda n veces. Si p es la probabilidad de cara, ¿cómo se distribuye la VA X: número de caras?

Distribución binomial
La VA X: "nº de éxitos en n-ensayos de Bernouilli" tiene **distribución binomial**:

$$X \sim \mathcal{B}(n,p) \iff P(X=x) = \mathcal{B}(x \mid n,p) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0, 1, \cdots, n.$$

El caso $n = 1$ suele llamarse **distribución de Bernouilli.**

Ejercicio: Esperanzas
Expresa X como $X = I_1 + I_2 + \cdots + I_n$, donde I_i es la función indicatriz del suceso "Éxito" en el lanzamiento i-ésimo, para demostrar que:

$$\mu = np \qquad \sigma^2 = npq \qquad q = 1 - p$$

R tiene comandos para trabajar con las distribuciones más famosas. Estos comandos comienzan con *d*, *p*, y *r*. Luego toman un sufijo que describe la distribución. Para la distribución Binomial:

Comando	¿Qué hace?
dbinom(k,n,p)	$P(X = k)$
pbinom(k,n,p)	$P(X \leq k)$
qbinom(probs,n,p)	Calcula los cuantiles especificados por las *probs*
rbinom(k,n,p)	Simula k VAs

Distribución geométrica

Ejercicio:
Se tira una moneda hasta que sale la primera cara. Si la probabilidad de cara es p, ¿cuál es la distribución de la VA X: "nº de cruces hasta la primera cara"?

Distribución geométrica
Cuenta el número de fallos X antes de obtener el primer éxito en una serie de experimentos de Bernouilli:

$$X \sim \mathrm{Geom}(p) \Longleftrightarrow P(X = x) = \mathrm{Geom}(x \mid p) = (1 - p)^x p.$$

La esperanza y la varianza son:

$$\mu = \frac{q}{p} \qquad \sigma^2 = \frac{q}{p^2} \qquad q = 1 - p.$$

En `R` tenemos las funciones *dgeom*, *pgeom*, ...

Danger!
A veces la distribución geométrica se define como el número de intentos X hasta obtener el primer éxito. En tal caso:

$$P(X = x) = (1 - p)^{x-1} p.$$

En este curso daremos prioridad a la otra parametrización por ser la que usa `R`.

Ejercicio:
Se tira una moneda con probabilidad de cara p. Sea la VA X: "nº de cruces hasta obtener 5 caras". ¿Cuál es la distribución de esta variable aleatoria?

Distribución binomial negativa
Cuenta el número de fallos hasta el r-ésimo éxito.

$$X \sim \text{NegBin}(r, p)$$

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

La esperanza y los momentos son:

$$\mu = \frac{rq}{p} \qquad \sigma^2 = \frac{rq}{p^2}.$$

En R tenemos los comandos *dnbinom*, *pnbinom*, *rnbinom*, ...

Ejercicio:
En cierto hospital, un 12 % de los pacientes no acude a su cita. Si un equipo médico es capaz de atender 100 personas en un día, cuál es la probabilidad de que una persona se quede sin atender si se dan 110 citas en un día. ¿Cuántas citas se pueden dar sin que dicha probabilidad exceda el 5 %?

Ejercicio:
De 2000 familias con 4 niños cada una, cuántos te esperarías que tuviesen a) al menos 1 niño (masculino) y b) 2 niñas.

Ejercicio:
Un enfermero necesita 10 radiografías de la pierna de un niño. Hay un 70 % de probabilidad de que el niño esté quieto durante la prueba. ¿Cuál es la probabilidad de que se necesiten más de 15 pruebas?

Distribuciones discretas

“Modificaciones” de la distribución binomial

Distribución hipergeométrica

La distribución binomial aparece en el muestreo con reemplazamiento, mientras que la **distribución hipergeométrica aparece en el muestreo sin reemplazamiento.**

Ejercicio:
Tenemos una caja con m bolas blancas y n bolas negras. Extraemos k de ellas sin reemplazamiento. ¿Cuál es la probabilidad de obtener x bolas blancas?

Distribución hipergeométrica
De una urna con m bolas blancas y n bolas negras extraemos k bolas sin reemplazamiento. Sea X: "número de bolas blancas extraídas". Entonces:

$$X \sim \text{HypGeom}(m, n, k)$$

$$P(X = x) = \text{HypGeom}(x \mid m, n, k) = \frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}} \qquad x = 0, 1, 2, \cdots$$

La media y la varianza son:

$$\mu = \frac{km}{m+n} \qquad \sigma^2 = \left(\frac{m+n-k}{m+n-1}\right) k \frac{\mu}{k}(1 - \frac{\mu}{k})$$

En R tenemos los comandos *dhyper*, *phyper*, *rhyper*, ...

La distribución binomial aparece en una sucesión de ensayos independientes, donde cada ensayo tiene dos posibles resultados. **La distribución multinomial surge también en una serie de ensayos independientes, pero cada ensayo tiene múltiples posibles resultados.**

Ejercicio:
Se tira un dado 12 veces. ¿Cuál es la probabilidad de obtener dos 1s, dos 2s, ... y dos 6s? ¿Y la distribución general para x_1 1s, x_2 2s, ...?

Distribución multinomial
Sea $\boldsymbol{X} = [X_1, X_2, \cdots, X_k]$ el vector que cuenta el número de éxitos para k-categorías excluyentes en n ensayos independientes. Si en cada ensayo las probabilidades de éxito para cada categoría son $\boldsymbol{p} = [p_1, p_2, ..., p_k]$, entonces:

$$\boldsymbol{X} \sim \text{MultiNom}(n, \boldsymbol{p}) \iff P(\boldsymbol{X} = \boldsymbol{x}) = \text{MultiNom}(\boldsymbol{x} \mid n, \boldsymbol{p}) = \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

Las esperanzas son

$$\mathbb{E}[X_i] = np_i \qquad \text{Var}[X_i] = np_i(1 - p_i).$$

Los comandos en R siguen la nomenclatura esperada: *rmultinom*, *dmultinom*, ...

Distribuciones discretas

Distribución de Poisson

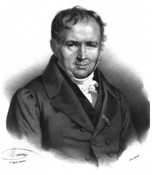
Distribución de Poisson

Una VA binomial/multinomial tiene prefijado el número máximo de casos de éxito. ¿Qué ocurre si **n no está acotado**?

Ejemplo: VAs discretas no acotadas

- El número de bebés nacidos en un hospital en una semana.
- El número de soldados muertos por una patada de caballo en un cuerpo de la caballería prusiana.
- El número de partículas emitidas en la desintegración de una partícula alpha.
- El número de accidentes en una carretera.

Si los eventos modelados ocurren de **forma independiente y pueden considerarse raros** (dos eventos no pueden ocurrir a la vez), la **distribución de Poisson** es la elección más habitual.



Distribución de Poisson

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

$$P(X = x) = \mathcal{P}(x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad x = 0, 1, 2, \cdots, \infty$$

Siméon Denis Poisson.

La media y varianza de la distribución de Poisson son:

$$\mu = \lambda \qquad \sigma^2 = \lambda.$$

En R tenemos *dpois, ppois, ...*

Ejercicio:
Un informático desea modelar el nº ataques informáticos que su página web sufre cada día. Recolecta datos durante 100 días, obteniendo la siguiente **tabla de frecuencias**:

nº de ataques/día	0	1	2	3
frecuencia	45	35	15	5

¿Probabilidad de que, en un día cualquiera, haya al menos 1 ataque informático?

Ejercicio:
Se tira una moneda hasta que salen 10 caras. A continuación, se meten 10 bolas blancas y tantas bolas negras como cruces hayan salido en una urna. Se extraen 5 bolas de la urna sin reemplazamiento. ¿Cuál es el número más probable de bolas blancas y cuál es su probabilidad?

Distribuciones continuas

Distribuciones continuas

La distribución uniforme

La distribución uniforme

Distribución uniforme

$$X \sim \mathcal{U}(a,b) \Longleftrightarrow f(x) = \mathcal{U}(x \mid a,b) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{otherwise} \end{cases},$$

La media y varianza son

$$\mu = \frac{1}{2}(a+b) \qquad \sigma^2 = \frac{1}{12}(b-a)^2.$$

Los comandos en `R` son *runif*, *dunif*, ...

Empleamos la distribución uniforme si todos los puntos son **equiprobables**.

Distribuciones continuas

La distribución exponencial

Distribución exponencial

$$X \sim \text{Expo}(\lambda) \iff f(x) = \text{Expo}(x \mid \lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

La media y varianza son

$$\mu = \frac{1}{\lambda} \qquad \sigma^2 = \frac{1}{\lambda^2}.$$

En R tenemos los comandos *dexp*, *pexp*, *rexp*, ... Una propiedad fundamental de la distribución exponencial es que **no tiene memoria**:

$$P(X < t_1 + t_2 \mid X > t_1) = P(X < t_2).$$

Esta propiedad hace que un uso habitual de la distribución exponencial sea modelar **tiempos de espera, tiempos hasta el primer fallo**, etc.

Distribuciones continuas

Distribución normal

Distribución normal o Gaussiana

La distribución normal o Gaussiana es, posiblemente, la más importante de todas las distribuciones.

De Moivre y Gauss.

Distribución Normal

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

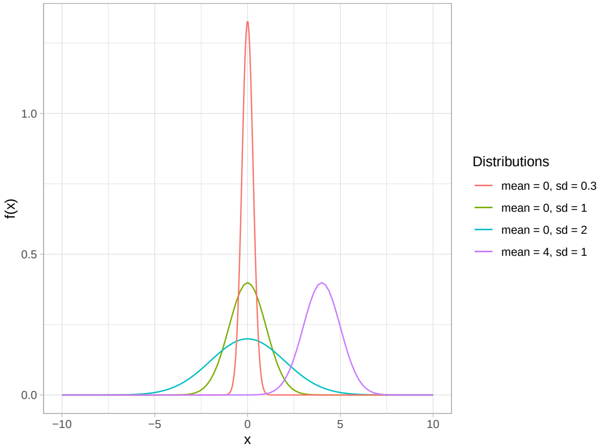
$$f(x) = \mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

La media y varianza de la Gaussiana son μ y σ^2.

En `R` tenemos las funciones *dnorm(x, mean, sd)*, *pnorm(q, mean, sd)*, ... ¡Ojo con sd!

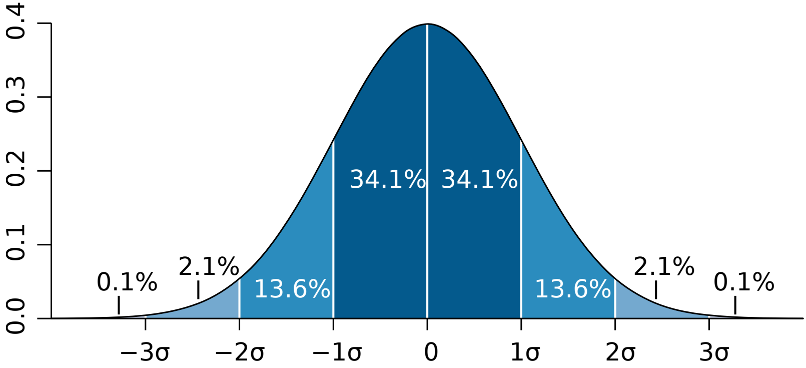
Ejercicio:
Prueba que la media y la varianza de $\mathcal{N}(\mu, \sigma^2)$ son μ y σ^2.

Distribución normal



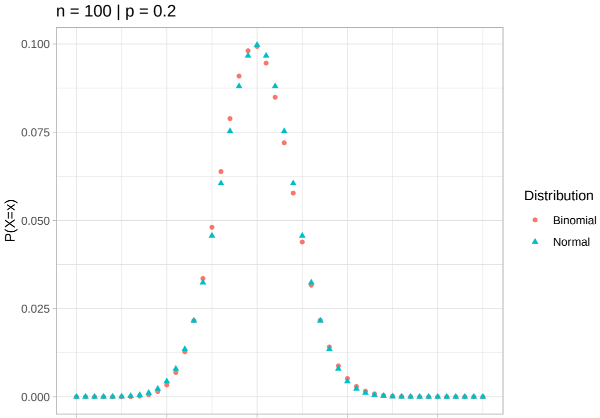
Independientemente de la media y varianza, siempre se cumple la

regla de 68-95-99.7

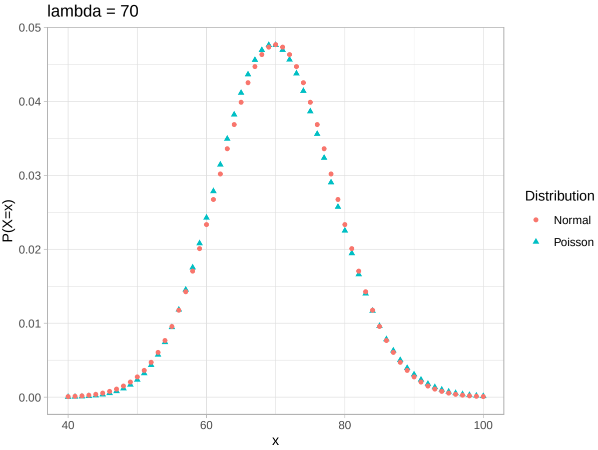


La regla de las 3 desviaciones estándares ya la usamos previamente en nuestro estudio de la distribución binomial. Pero, ¿qué relación guardan la distribución binomial y la distribución normal?

La respuesta es el **teorema del límite central**, que afirma que la **suma de muchas variables aleatorias independientes tiende a parecerse a una normal** (y recuerda que $X = I_1 + I_2 + \cdots + I_n$.)



La aproximación funciona también para VAs de Poisson (¿Por qué?)



Teorema del límite central (TCL)

De hecho, el teorema del límite central nos permite precisar cuál es la media y varianza de la Gaussiana límite.

Teorema del límite central
Si X_1, $X_2, \cdots, X_n$ son VAs independientes e identicamente distribuidas con media μ y varianza σ^2 finitas. Entonces $S_n = \sum_{i=1}^{n} X_n$ verifica

$$S_n \Rightarrow \mathcal{N}(n\mu, n\sigma^2)$$

Ejercicio:
Usa el TCL para demostrar que una distribución Binomial tiende a una distribución

$$\mathcal{N}(np, \sigma^2 = npq)$$

si n es grande. Dibuja ambas distribuciones para visualizar la coincidencia.

Ejercicio:
Usa el TCL para demostrar que una distribución de Poisson tiende a una distribución

$$\mathcal{N}(\lambda, \sigma^2 = \lambda)$$

si λ es grande. Dibuja ambas distribuciones para visualizar la coincidencia.

Suma de Normales independientes

Si el **número de variables sumadas es pequeño**, en general, **el TCL no funciona**. La única excepción es la siguiente:

Propiedad reproductiva de Normales independientes
Si X_1 y X_2 son VAs Normales independientes:

$$X_1 + X2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

Ejercicio:
Tres hermanos van a comer a casa de su abuelita. El 99.7 % de las veces, cada hermano come una cantidad de comida comprendida entre 1.2 y 1.8 Kg. ¿Cuál es la probabilidad de que logren acabar los 5 Kg de carne que su abuela ha preparado?

Ejercicio:
Has programado un robot asesino para acabar con tu profesor de estadística. El robot dispara al centro de su frío corazón, pero comete un error aproximadamente normal en cada una de las coordenadas x e y. La media de ambas normales es 0 y tiene desviación típica 5 cm. ¿Cuál es la probabilidad de que el disparo acabe a menos de 1 cm del centro del corazón? Resuelve por simulaciones.

Distribuciones mixtas

Ejercicio:
Supón que administras un centro de llamadas para una empresa de soporte técnico. La empresa recibe llamadas de clientes que buscan ayuda con tres tipos diferentes de problemas técnicos: problemas de software (S), problemas de hardware (H) y problemas de red (N). Las probabilidades de cada tipo de llamada son:

- La probabilidad de recibir una llamada tipo S es 0.4.
- La probabilidad de recibir una llamada tipo H es 0.3.
- La probabilidad de recibir una llamada tipo N es 0.3.

La duración de cada llamada (en minutos) es una variable aleatoria continua:

- Si la llamada es de tipo S, la distribución es exponencial con una media de 12 minutos.
- Si la llamada es de tipo H, la distribución es exponencial con una media de 10 minutos.
- Si la llamada es de tipo N, la distribución es Gaussiana con media de 9 minutos y desviación estándar de 2 minutos.

Si una llamada dura menos de 11 minutos, ¿cuál es el tipo de llamada más probable?