Prácticas de Estadística

Práctica 3: Variables aleatorias. Simulación.

Introducción:

La simulación de un sistema real se usa para prever resultados, comparar estrategias,... a un costo muy inferior a hacerlo realmente. Tradicionalmente se han desarrollado prototipos a escala, ensayos en condiciones controladas y seguras, etc. Actualmente, la simulación mediante ordenador resulta generalmente la más barata y es la más extensamente aplicada. (Leer 'Simulación' en Wikipedia).

Herramientas para simulación:

MATLAB básico dispone de las funciones **rand(m,n)** y **randn(m,n)** que devuelven una matriz de m filas y n columnas de números aleatorios siguiendo una Uniforme [0,1) y una N(0,1), respectivamente. Con ellas pueden simularse otras muchas, por ejemplo:

Uniforme [a,b]: Mediante **x=a+(b-a)*rand(m,n)**

Uniforme discreta en $\{Nmin, Nmin+1, \ldots, Nmax\}$**:** Mediante **x=Nmin+floor(Nmax*rand(m,n))** (las versiones nuevas de MATLAB disponen de **randi([Nmin,Nmax],m,n)** para este fin).

Normal(a,b): Mediante **x=a+b*randn(m,n)**

Además de estas, la toolbox "stats", simula fácilmente muchas otras funciones útiles:

Distribución	Función de densidad (pdf)	Función de Distribución (cdf)	Función cuantil (inv)	Números aleatorios (rnd)
Beta	betapdf	betacdf	betainv	betarnd
Binomial	binopdf	binocdf	binoinv	binornd
Binomial negativa	nbinpdf	nbincdf	nbininv	nbinrnd
χ^2	chi2pdf	chi2cdf	chi2inv	chi2rnd
Exponencial	exppdf	expcdf	expinv	exprnd
F Fisher-Sn.	fpdf	fcdf	finv	frnd
Gamma	gampdf	gamcdf	gaminv	gamrnd
Geométrica	geopdf	geocdf	geoinv	geornd
Hipergeométrica	hygepdf	hygecdf	hygeinv	hygernd
Log-normal	lognpdf	logncdf	logninv	lognrnd
Multinomial	mnpdf	mncdf	mninv	mnrnd
Normal multivariante	mvnpdf	mvncdf	-	mnrnd
Normal	normpdf	normcdf	norminv	normrnd
Poisson	poisspdf	poisscdf	poissinv	poissrnd
Rayleigh	raylpdf	raylcdf	raylinv	raylrnd
T-Student	tpdf	tcdf	tinv	trnd
Uniforme discreta	unidpdf	unidcdf	unidinv	unidrnd
Uniforme cont.	unifpdf	unifcdf	unifinv	unifrnd
Weibull	wblpdf	wblcdf	wblinv	wblrnd

Ejercicio 5:

Hallar una matriz de tamaño 5×10, con números aleatorios siguiendo la distribución especificada:

1. a) $\xi \rightsquigarrow N(4,7)$, b) $\eta \rightsquigarrow N(-3,2)$, c) $\zeta \rightsquigarrow t_{15}$ (T-Student).
2. a) Uniforme discreta en $\{1, 2, \ldots, 6\}$, b) Uniforme continua en [-2,2].
3. a) Binomial(100,0.4), b) Binomial(500,0.003).
4. a) Chi-Cuadrado con 20 g.d.l. b) Poisson con $\lambda = 9$, c) Rayleigh R(4).

Prácticas de Estadística

Práctica 4: Método de Montecarlo.

Introducción:

Consiste en estimar la probabilidad de que ocurra un suceso mediante la repetición del experimento aleatorio un gran número de veces. Se basa en las **"Leyes de los grandes números"** aplicada al experimento de Bernouilli.

Leyes de los grandes números

Si realizamos N experimentos aleatorios de forma independiente de una misma distribución ξ, entonces la media de los resultados obtenidos $\bar{x} = \frac{\sum_i^N x_i}{N}$ converge a la media de ξ. Esto es: $\lim_{\mathbf{N}\to\infty} \bar{\mathbf{x}} = \mu = \mathbf{E}(\xi)$

Así, la media de las N realizaciones de un experimento de Bernouilli ($\hat{p} = \frac{N_a}{N}$, N_a=Número de veces en que ocurre A), converge a la media de la Bernouilli, es decir p. Además se sabe que para N grande ($N > 30$), resulta: $\hat{p} \rightsquigarrow N\left(p, \sqrt{\frac{pq}{N}}\right)$, lo que nos lleva a:

$$P\left(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{N}} \leq p \leq \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{N}}\right) = \alpha \quad \Rightarrow \quad I_p = \left[\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{N}},\ \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{N}}\right]$$

Otra convergencia importante es que la media de N variables aleatorias iguales e independientes de media μ y desviación típica σ, cuando N es grande, converge a la media de la población según una normal: $\bar{x} \rightsquigarrow N\left(\mu, \frac{\sigma}{\sqrt{N}}\right)$, por tanto:

$$P\left(\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{N}} \leq \mu \leq \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{N}}\right) = \alpha \quad \Rightarrow \quad I_\mu = \left[\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{N}},\ \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{N}}\right]$$

Ejercicio 6:

Supongamos la región rectangular R delimitada por los puntos $(0,0), (20,0), (20,10)$ y $(0,10)$, donde las unidades expresan Km. Tenemos antenas de TV situadas en los puntos $(-1,5)$, $(5,4)$, $(15,7)$, $(10,1)$, $(12,2)$, $(19,1)$ y $(15,2)$, que pueden recoger/enviar la señal a algún punto de R y que tienen un alcance de 3Km.

Si todos los puntos de R tienen la misma probabilidad de emitir una señal, hallar mediante simulación con 10.000 ejecuciones:

1. La proporción de señales emitidas desde R que serán recogidas por alguna antena y el área de la misma.
2. La proporción de señales emitidas desde R que serán recogidas por varias antenas.
3. Si queremos seleccionar el lugar idóneo para una nueva antena entre $A = (1,1)$, $B = (10,6)$ y $C = (14,4)$. En el sentido que incrementa más la región con cobertura, ¿Cuál es mejor?

Solución:

a: Podemos programarlo realizando las experiencias de forma sucesiva y empleando contadores de veces en que ocurre el experimento, o realizando las 10.000 ejecuciones a la vez (más rápido, pero ocupa más memoria).

De todas formas, vamos a hacerlo de forma simultanea:

a1) Primero debemos generar 10.000 veces una posición aleatoria en $[0,20] \times [0,10]$:

Lo hacemos con **N=10000; x=[20*rand(N,1) 10*rand(N,1)];**

Tendremos una matriz 10000×2 donde cada fila es un punto aleatorio en R.

a2) Estudiamos si el punto (fila de x) está en la cobertura de cada una de las antenas:

Para antena en [-1,5]: **xx=[x(:,1)+1 x(:,2)-5];d1=sqrt(xx(:,1).^2+xx(:,2).^2);c1=(d1<=3);**

Para antena en [5,4]: **xx=[x(:,1)-5 x(:,2)-4];d2=sqrt(xx(:,1).^2+xx(:,2).^2);c2=(d2<=3);**

...

Para antena en [15,2]: **xx=[x(:,1)-15 x(:,2)-2];d7=sqrt(xx(:,1).^2+xx(:,2).^2);c7=(d7<=3);**

a3) Calculamos si tiene alguna cobertura, lo que podemos hacer mediante $c = c1|c2|c3|c4|c5|c6|c7$ pero teniendo en cuenta el apartado b, resulta mejor:

c=c1+c2+c3+c4+c5+c6+c7; que nos dice el número de antenas en su entorno.

a4) Calculamos el número de puntos con cobertura: **Na=sum(**$c > 0$**)**

a5) Estimamos la proporción y el área con cobertura: **p=Na/N, S=p*20*10, q=1-p**

a6) Intervalo de confianza para p y S al 95% ($z_{\frac{\alpha}{2}} = 1.96$ **z=norminv(0.975,0,1)**):

Ip=[p-1.96*sqrt(p*q/N), p+1.96*sqrt(p*q/N)], Is=Ip*200

b) Ya está hecho todo el trabajo hasta calcular Na, falta:

Nb=sum($c > 1$**)**
b1) Estimamos la proporción y el área con cobertura: **pb=Nb/N, S=pb*20*10,qb=1-pb**
b2) Intervalo de confianza para pb y Sb al 95%:
Ipb=[pb-1.96*sqrt(pb*qb/N), pb+1.96*sqrt(pb*qb/N)], Isb=Ipb*200

c) Resulta repetitivo, pues basta con añadir una antena en el apartado a) y ver cuál da un $\hat{p}$ mayor, pero sirve para ver como la simulación permite evaluar estrategias.

Ejercicio 7:

Repetir el ejercicio anterior, pero ahora se va a tener en cuenta la existencia de un núcleo de población en el punto (12,6) y una carretera, por lo que los puntos aleatorios $(x, y) \in R$ deberán seleccionarse considerando este hecho desde una normal bivariante de media (12,6) y matriz de covarianzas $V = \begin{pmatrix} 3 & 0.5 \\ 0.5 & 2 \end{pmatrix}$. Se considera una población de 25000 personas. Hallar la proporción con cobertura y el número de las mismas.

Nota: Ahora un punto (x,y) aleatorio, podrá salirse de R y debemos filtrar los válidos.

Solución

a1) Mediante **help stats** y **help mvnrnd** vemos que podemos generar los 10000 vectores aleatorios con:

N=10000; x=mvnrnd([12 6],[3 0.5;0.5 2],N);

a1-bis) Tenemos que ver si el punto es de R:

R=(x(:,1)<=20).*(x(:,1)>=0).*(x(:,2)>=0).*(x(:,2)<=10);

Nr=sum(R>0);

los puntos a2) y a3) serían iguales, pero a4) y siguientes quedarían:

Na=sum(c.*R>0);

p=Na/Nr, q=1-p, Ip=[p-1.96*sqrt(p*q/Nr), p+1.96*sqrt(p*q/Nr)]

Pobl=25000*p, Ipobl=25000*Ip