Práctica 2. Números y variables aleatorias

OBJETIVOS DE LA PRÁCTICA

1. Realizar varios contrastes empíricos sobre la bondad de ajuste de generadores de números aleatorios
2. Analizar la aleatoriedad de un conjunto de datos

FORMA DE REALIZAR LA PRÁCTICA:

Se trabajará en el entorno MATLAB, desarrollando códigos que ofrezcan soluciones a los problemas planteados.

TRABAJO A PRESENTAR:

Se debe entregar un fichero comprimido que contenga

- El código fuente de los programas elaborados con los comentarios internos necesarios para su identificación y seguimiento.
- Una memoria en MS-WORD con el contenido siguiente:
    - Solución para cada ejercicio y apartado.
    - Las tablas, matrices y gráficos generados por la ejecución del programa, así como las respuestas razonadas a las preguntas de cada enunciado.

Introducción.

Se quiere realizar una simulación de algoritmos de planificación de trayectorias sobre un robot industrial utilizando generadores de números aleatorios. Se pretende que el robot vaya de una configuración inicial a otra final siguiendo un conjunto de configuraciones intermedias. Para ello es necesario crear inicialmente un amplio conjunto de configuraciones aleatorias para posteriormente generar un camino desarrollando un algoritmo de búsqueda.

Una configuración aleatoria de robot esta compuesta por un vector de valores que se corresponden con el número de grados de libertad (seis en nuestro caso). Es decir, tenemos que obtener seis valores aleatorios para cada configuración de robot. Dichos valores deben estar comprendidos entre los valores mínimo y máximo para cada articulación según la tabla adjunta.

Número Articulación	Mínimo	Máximo
1	-160	160
2	-137.5	137.5
3	-142.5	142.5
4	-270	270
5	-112.5	112.5
6	-270	270

La generación de estas configuraciones aleatorias se pretende realizar con generadores implementados en ciertas librerías (en nuestro caso en el entorno de Matlab) en donde se puede imponer el tipo de distribución por la que se quieren generar los datos aleatorios.

Para simplificar la explicación, supongamos una única articulación del robot. En nuestra aplicación, una distribución de tipo uniforme U(a,b) supondría proporcionar igual peso a cada valor aleatorio generado (posición articular) mientras que el uso de una distribución normal $N(\mu,\sigma)$ presupone que los valores aleatorios generados son más probables alrededor de un valor - cercano a media de la distribución - que en otros.

El objetivo de esta práctica es analizar y comprobar el grado de fiabilidad de generadores de números aleatorios así como medir la aleatoriedad de los datos obtenidos para realizar aplicaciones de simulación como la descrita. A partir de ahora trabajaremos en un intervalo normalizado [0,1] de valores de cada articulación.

Ejercicio 2.1

Enunciado

En este ejercicio se pretende comprobar si los generadores de números aleatorios comerciales que incluye MATLAB, correspondientes a varias distribuciones (Normal, Uniforme y Binomial (ver apartado 3)) pasan determinados contrastes de bondad de ajuste. Las distribuciones Uniforme y Normal son continuas, mientras que la distribución Binomial es discreta. Para ello, se realizarán los contrastes de hipótesis χ^2

y Kolmogorov-Smirnov (K-S) para cada una de las tres distribuciones. Se tomará un tamaño de muestra de 500 valores y 12 clases (para las distribuciones continuas).

Documentación

1. Test de hipótesis de χ^2.

Se realiza para variables discretas y continuas agrupadas en clases. La expresión de cálculo del estadístico χ^2 se realiza a través de la expresión:

$$\chi^2 = \sum_{nc} \frac{(f_a - ft_a)^2}{ft_a}$$

En esta expresión f_a (apartado 4) corresponde a las frecuencias absolutas recogidas en el experimento mientras que f_{ta} (apartado 5) son las correspondientes frecuencias teóricas. El p_valor para un nivel de confianza del 95% se obtiene a través de la función chi2inv de la siguiente manera:

P_valor=chi2inv(0.95,*nc-1*)

donde *nc* es el número de clases.

Si χ^2<*P_valor* se acepta la hipótesis nula por lo que se considera que los datos aleatorios generados se consideran que se ajustan a la distribución teórica con un nivel de confianza del 95%.

En esta práctica consideramos que el tamaño de la muestra es *n*=500 y el numero de clases *nc*=12.

Realizar los test de χ^2 para generadores de números aleatorios de distribuciones N(0,1), U(0,1), B(10, 0.2).

2. Test de hipótesis de K-S

Se realiza para variables continuas agrupadas o no agrupadas en clases. El cálculo del estadístico de K-S se calcula a través de la expresión:

$$D_n = \sup_x |F_a(x) - F_t(x)|$$

donde $F_a(x)$ (ver apartado 6) es la frecuencia relativa acumulada sobre los datos ordenados hasta el dato *x* y $F_t(x)$ (ver apartado 7) es la frecuencia relativa teórica acumulada hasta el dato *x*. El valor de D_n está tabulada para valores grandes y pequeños de n (n<40). Sin embargo, en esta práctica se realizará el contraste K-S utilizando directamente la función:

H=kstest(*datos,Matriz_cdf*)

donde *Matriz_cdf* es una matriz de dos columnas. En la primera columna están los datos ordenados recogidos en el experimento (datos empíricos), mientras que en la segunda columna están las frecuencias relativas acumuladas ordenadas teóricas *Ft.*

Si *H*=0 entonces se acepta la hipótesis, por lo que se considera que los datos generados aleatoriamente se ajustan a una determina distribución.

Realizar los test de Kolmogorov-Smirnov (K-S) para generadores de números aleatorios de distribuciones N(0,1) y U(0,1).

3. Los generadores de números aleatorios que se tomarán en la práctica se obtienen a través de las siguientes funciones:

1) *datos*=randn(1,n). Genera un vector de números aleatorios (de dimensión n) siguiendo una distribución Normal de media 0 y desviación típica 1, N(0,1). En la distribución normal N(μ,σ), μ es la media de la distribución y σ la desviación típica.

2) *datos*=rand(1,n). Genera un vector de números aleatorios (de dimensión n) comprendidos entre 0 y 1 siguiendo una distribución Uniforme (U(0,1)). En la distribución uniforme N(a,b), a y b son los extremos del intervalo.

3) la función randtool aplicado a distribución binomial con parámetros (10,0.2) proporciona la variable *binorv* (*datos=binorv'*) que contiene el conjunto de números aleatorios buscado. En la distribución binomial B(s,p), s es el número de experimentos y p es la probabilidad de obtener éxito en un experimento.

En este caso, antes de realizar el test de χ^2 debes generar y guardar la variable *binorv* siguiendo los siguientes pasos:

- Ejecuta randtool desde la línea de comandos.
- Selecciona la distribución Binomial.
- Poner 500 muestras (samples), 10 pruebas (trials) y probabilidad de resultados (Prob) de 0.2.
- Exportar al espacio de trabajo en la variable *binorv*
- Guardar en un archivo MATLAB ; save *path**nombre_archivo binorv*

Para utilizar estos datos aleatorios en el código de programa

- Carga los valores aleatorios , load *path**nombre_archivo*
- Asigna la variable datos: *datos*=binorv'

Para cada caso 1), 2) y 3) realizar una salida gráfica mostrando los conjuntos de valores aleatorios generados.

4. Calculo de f_a : Frecuencias absolutas **reales** recogidas en el experimento

Como se ha dicho, las frecuencias absolutas son utilizadas en el test de χ^2.

En los casos de las distribuciones continuas (N y U), se debe realizar una agrupación en clases de los datos aleatorios (variable *datos*) y obtener las frecuencias absolutas a través de la función hist mediante:

[f_a,*posición_centros*]=hist(*datos, numero_clases*)

donde f_a es la frecuencia absoluta y *posicion_centros* son los centros de las clases. Calcular las posiciones de los extremos de cada clase y guardar en un vector *posición_extremos*

En el caso de la distribución discreta binomial B, no cabe realizar agrupaciones en clases siendo:
numero_clases=length(min(*datos*):max(*datos*))
posicion_centros= min(*datos*):max(*datos*)
En este caso, se utiliza igualmente la función hist para calcular las frecuencias absolutas:

f_a=hist(*datos, numero_clases*)

Para cada caso realizar una salida gráfica mostrando el histograma de frecuencias absolutas.

5. Cálculo de f_{ta}: Frecuencias **teóricas** absolutas

En los contrastes de hipótesis se pretende contrastar que los datos obtenidos con los procedimientos anteriores corresponden realmente a una muestra aleatoria teórica que sigue la distribución determinada. El cálculo de los valores teóricos de las frecuencias absolutas para cada caso de distribución, se realiza calculando la probabilidad teórica y multiplicando por el tamaño de la muestra. Así $ft_a=pt*n$. donde la probabilidad teórica puede ser calculada para las distribuciones escogidas como sigue:

1) Distribución Normal. La probabilidad teórica se calcula a través de las funciones diff y normcdf como: *pt*=diff(normcdf(*posición_extremos,0,1*)). La variable *posición_extremos* debe haber sido calculada en el apartado 4. Esta operación calcula el área encerrada por la distribución para cada intervalo que es justamente la probabilidad teórica.

2) Distribución Uniforme. La probabilidad teórica se calcula a través de las funciones diff y unifcdf como: *pt*=diff(unifcdf(*posición_extremos,0,1*))

3) Distribución Binomial. La probabilidad teórica se calcula a través de la función binopdf como: *pt*=binopdf(*posicion_centros,10,0.2*). En este caso, ya que la distribución es discreta no se calcula en área sino que se recoge la función de densidad de probabilidad para los valores *posicion_centros*. Tener en cuenta que *posicion_centros* deben ser valores enteros.

Para cada caso realizar una salida gráfica mostrando las funciones de probabilidad reales (f_a/n) y teóricas (ft_a/n) superpuestas en la misma gráfica.

6. Cálculo de $F_a(x)$. Frecuencias relativas acumuladas de los datos **reales** ordenados

Como se ha dicho, las frecuencias empíricas relativas acumuladas son utilizadas para el test de K-S. Ya que la frecuencia de cada dato es 1/n (en caso de datos no agrupados), la frecuencia acumulada del i-ésimo dato es $F_a(x_i)=i/n$ y la frecuencia del último dato $F_a(x_n)=1$.

7. Cálculo de $F_t(x)$. Frecuencias relativas acumuladas **teóricas** de los datos ordenados

El cálculo de los valores teóricos de las frecuencias relativas acumuladas para cada caso de distribución se realiza:

1) Distribución Normal. Utiliza las funciones normcdf y sort obteniendo *Ft*=normcdf(sort(*datos*),*0,1*). La función normcdf corresponde a la función de distribución normal acumulada.
2) Distribución Uniforme. Utiliza las funciones unifcdf y sort obteniendo *Ft*=unifcdf(sort(*datos*),*0,1*). La función unifcdf corresponde a la función de distribución uniforme acumulada.
3) Distribución Binomial. Utiliza las funcion binocdf obteniendo *Ft*= binocdf(sort(*datos*),*10,0.2*). La función binocdf corresponde a la función de distribución binomial acumulada.

Para cada caso realizar una salida gráfica mostrando las frecuencias relativas acumuladas reales (F_a) y teóricas (Ft) superpuestas en la misma gráfica.

Ejercicio 2.2

Enunciado

1. Generar un test de números aleatorios de una distribución normal mediante la función randn *y medir el contraste de bondad de ajuste para los test de Lilliefors, Bera-Jarque y KS. Se tomará un tamaño de muestra de 500 valores y 12 clases.*

Utilizar la función normplot*(datos) para realizar un test gráfico de normalidad. Observa que si los puntos se ajustan muy bien a la línea recta es síntoma de que se acepta la bondad de ajuste*

2. Realizar 800 pruebas de los test referidos y obtener el porcentaje de tests que son rechazados para cada uno de los tres casos. Mostrar gráficamente con normplot *los casos en los que haya a la vez un rechazo de bondad de ajuste para los tres tests .*

Documentación

Se pretende comprobar cuál es el porcentaje de aceptación de contraste de hipótesis para 800 test utilizando los test de Lilliefors, Bera-Jarque y KS sobre distribuciones normales.

Estos tests están implementados directamente sobre los datos del experimento por las funciones: lillietest, jbtest y kstest. Al igual que antes tan sólo hay que recoger el valor H1= lilliestest(*datos*), H2= jbtest (*datos*), H3= kstest (*datos*) y comprobar que el valor devuelto es 0 ó 1. En caso de obtener valor 1 la hipótesis de normalidad es rechazada. Los datos son generados como en el apartado anterior mediante la funcion randn.

Ejercicio 2.3

- Test de rachas de crecimiento/decrecimiento.

Enunciado

Generar una sucesión de números aleatorios utilizando 1) un generador para distribuciones normales y 2) un generador para distribuciones uniformes y realizar el contraste de aleatoriedad de rachas de crecimiento/decrecimiento.

Documentación

El test de contraste de rachas de crecimiento/decrecimiento tiene por objetivo determinar si los datos obtenidos en un experimento son aleatorios o no. El test se realizará para dos sucesiones de observaciones correspondientes a los generadores de números aleatorios de distribuciones normal randn(1,n) e uniforme rand(1,n), tomando n=500.

Dada una sucesión de observaciones X_1,X_2,....X_n (en nuestro casos corresponde a la variable 'datos') construir una sucesión binaria siendo 1 si $X_i<X_{i+1}$ y 0 si $X_i>X_{i+1}$. Encontrar el número de rachas de ceros y unos en la sucesión binaria y contabilizar el total de rachas T.

Para cada caso realizar una salida gráfica mostrando los conjuntos de valores aleatorios generados.

Bajo la hipótesis nula de aleatoriedad se supone que *T* es una variable aleatoria con distribución normal: $N(\frac{2n-1}{3},\frac{16n-29}{90})$, siendo *n* el numero de muestras de la sucesión, $\frac{2n-1}{3}$ la media y $\frac{16n-29}{90}$ la desviación estándar.

Obtener los dos valores de aceptación (*p_valores*) de la hipótesis trabajando con un 0.05% de significación bilateral utilizando la función norminv:

p_valores=norminv([0.025,0.975], *media*, *desviación_estandar*)

si el valor de *T* está en el intervalo [p_valores(1), p_valores(2)] se acepta el contraste de aleatoriedad de rachas.

Mostrar el valor de T calculado, los dos p_valores y el resultado final del test.

- Test de rachas de mediana

Enunciado

Generar una sucesión de números aleatorios utilizando 1) generador para distribuciones normales y 2) generador para distribuciones uniformes y realizar el contraste de aleatoriedad de rachas de mediana.

Documentación

El test de contraste de rachas de mediana tiene por objetivo determinar si los datos obtenidos en un experimento son aleatorios o no. El test se realizará para dos sucesiones de observaciones a través de los generadores de números aleatorios de distribuciones normal randn(1,n) e uniforme rand(1,n), tomando n=500.

El procedimiento es similar al anterior pero en este caso se tiene en cuenta si los datos está por encima o por debajo de la mediana de la sucesión de valores empíricos. El valor de la mediana se calcula a través de la función median(*datos*). Suponiendo que n_1 y n_2 son los números de datos que están por encima y por debajo de la mediana el número total de rachas T tiene teóricamente una distribución normal:

$$N\left(\frac{2n_1n_2}{n_1+n_2}+1,\left[\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}\right]^{1/2}\right)$$

En el caso de que n_1 y n_2 sean grandes, $n_1 \approx n_2 \approx \frac{n}{2}$ la distribución se puede aproximar a:

$$N\left(\frac{n}{2},\frac{n}{4}\right)$$

Igual que en caso anterior, si T es el número total de rachas encontradas, utilizar

p_valores=norminv([0.025,0.975], *media*, *desviación_estandar*)

para verificar el contraste de aleatoriedad comparando T con los valores p_valores(1) y p_valores(2). Incluir los dos casos expuestos para $N\left(\frac{2n_1n_2}{n_1+n_2}+1,\left[\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}\right]^{1/2}\right)$ y $N\left(\frac{n}{2},\frac{n}{4}\right)$.

Mostrar el valor de T calculado, los dos p_valores y el resultado final del test.