PRÁCTICA 5. LA DISTRIBUCIÓN NORMAL

Objetivo

El objetivo de la presente práctica informática es aplicar los conceptos vistos en clase sobre la distribución Normal. Comprobar sus propiedades y compararla con otras distribuciones, utilizando para ello representaciones gráficas y parámetros, algunos de los cuales ya se introdujeron en la Unidad Didáctica 2.

1. Características de la distribución Normal

a) Observa la forma de la función de densidad f(x) que caracteriza a la distribución Normal de la variable aleatoria DUREZA DE LOS ASIENTOS utilizada en el Ejercicio 16 de la UD4-Parte 3.

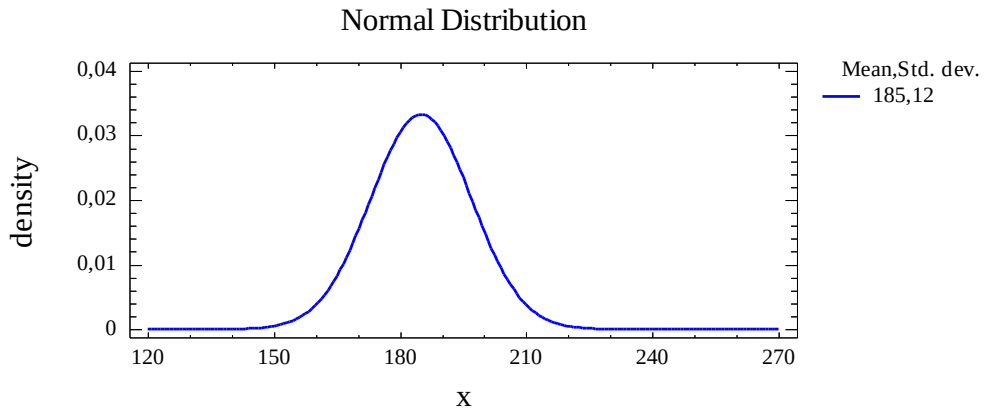
Opciones: ***Plot → Probability Distributions (Normal)***

Con la tecla derecha del ratón abrir la ventana en la que se selecciona

Analysis Options: mean=185 Nw, Std. Dev.=12 Nw

Respuesta:

En la ventana superior derecha aparece el gráfico de la función de densidad de la Normal de media 185 y desviación típica 12. Tiene forma simétrica y encierra un área con forma de campana centrada en el valor 185.



b) Calcula un intervalo de valores de DUREZA que contenga el 95% de TODOS los asientos fabricados (Población). Compara el resultado con las propiedades enunciadas en el apartado 1.5.2 de la UD4-Parte 3.

Opciones: ***Tabular Options → Inverse CDF***

Y con la tecla derecha del ratón abrir la ventana para seleccionar **Pane Options** y en CDF poner 0,025 y 0,975

Respuesta:

Se estudió en las propiedades de la distribución normal que

$P(m-2\sigma<N(m,\sigma)<m+2\sigma)\approx 0{,}95$

Por tanto la probabilidad acumulada por debajo de m-2σ es aproximadamente 0,025 y por debajo de m+2σ es aproximadamente 0,975. Al haber indicado esas dos probabilidades en CDF el programa nos dará los valores 161,48 y 208,52 que están muy próximos a 185-2.12=161 y 185+2.12=208.

2. Comparación de la distribución Normal con otras distribuciones

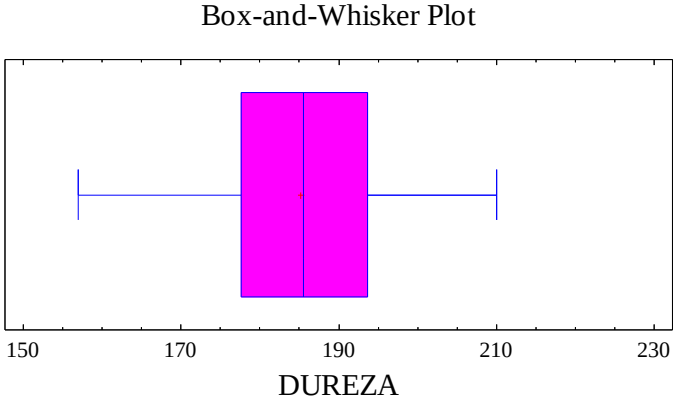
Se ha tomado una muestra de 100 asientos y se ha medido su dureza (los valores se recogen en la variable DUREZA). Asimismo, se ha tomado una muestra de 100 pantallas LCD (Ejercicio 12 de la UD4-Parte 2) y se ha medido el tiempo (horas) hasta que dejan de funcionar correctamente (los valores se recogen en la variable TIEMPO). Los datos de las variables DUREZA y TIEMPO se encuentran en el fichero **PRACT5-GII.SF3** disponible en PoliformaT. Descargar el fichero de Recursos..practicas….ficheros de datos. Abrirlo con **File…Open…Open Data File**. En buscar en poner el subdirectorio en el que se haya descargado el fichero.

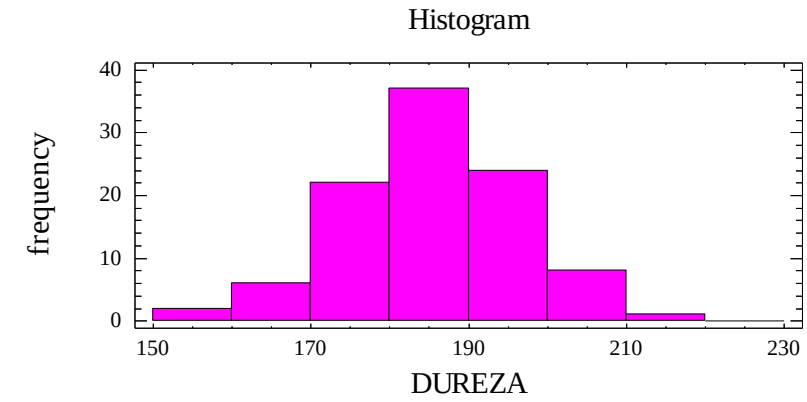
a) Construye para cada una de las muestras un histograma, el diagrama *Box&Whisker*, representa los valores en Papel Probabilístico Normal y compáralos. ¿Qué puedes decir respecto de las distribuciones de las cuales provienen?

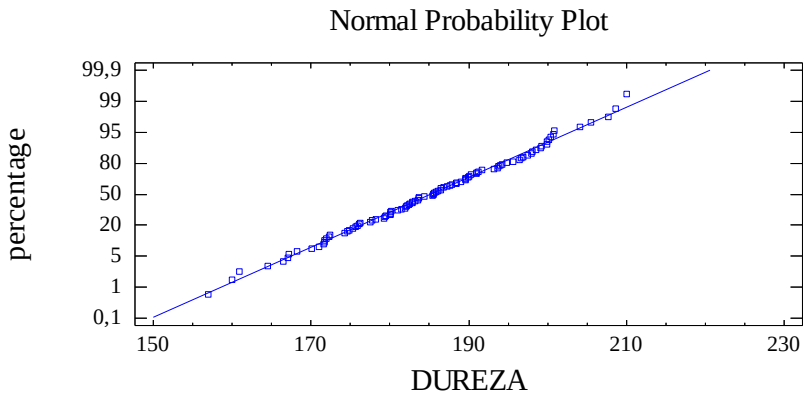
Opciones: **Describe…Numeric Data…One Variable Analysis**. Primero ponemos en **Data** La variable DUREZA. Se abrirá la ventana de análisis con el diagrama *Box&Whisker*. Para obtener el Histograma y la representación en Papel Probabilístico Normal seleccionar el icono **Graphical Options** y en la ventana que se abre marcar **Frequency Histogram** y **Normal Probability Plot**. Repetir la operación con la variable TIEMPO.

Respuesta:

Los tres gráficos para la variable DUREZA son:

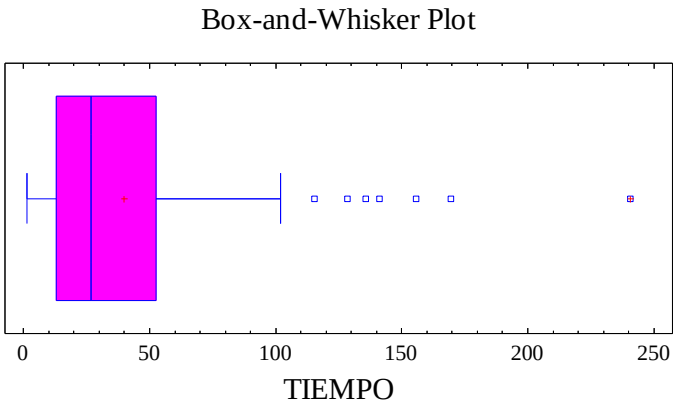


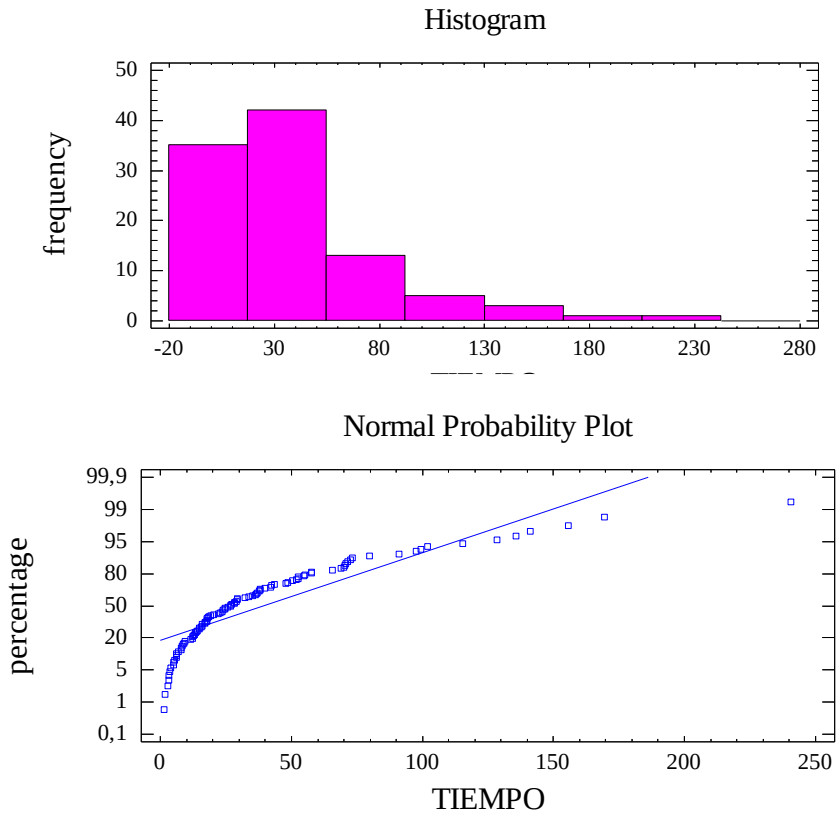




Se observa en el diagrama Bos&Whisker que la distribución es simétrica. En el histograma la distribución de frecuencias tiene forma aproximada de campana. En el papel probabilístico normal recordad que se representan en el eje Y los porcentajes acumulados y en el eje X los valores ordenados de la muestra. La escala del eej Y es especial para la distribución Normal, de forma que si los daos siguen esta distribución al hacer la representación aparecen cercanos a una línea recta. Esto es lo que se observa para la variable DUREZA. Por tanto la conclusión es que la DUREZA sigue distribución Normal.

Para la variable TIEMPO los tres gráficos son:





Se observa en el diagrama Box&Whisker de TIEMPO que la variable es asimétrica positiva. El histograma confirma con la distribución de frecuencias esa distribución. Finalmente la representación en papel probabilístico normal es una curva. Por tanto el TIEMPO no sigue una distribución normal.

b) Obtener la media, la mediana, la desviación típica y los parámetros estándar de asimetría y curtosis para las variables TIEMPO y DUREZA. ¿Qué se observa entre la media y la mediana de cada variable? ¿Qué puedes decir respecto de la asimetría y curtosis?

Para obtener estos parámetros muestrales desde la misma opción **Describe…Numeric Data…One Variable Analysis** hay que mirar la venta que tiene como título **Summary Statistics** y con la tecla derecha del ratón abrir una ventana en la que seleccionados **Pane Options** y en ella marcamos **Average Median Std.Deviation Std Skewness y Std Kurtosis**. Lo hacemos para cada una de las variables.

Respuesta:

Los parámetros pedidos para DUREZA son:

Summary Statistics for DUREZA

Count = 100
Average = 185,2
Median = 185,484
Standard deviation = 11,3074
Stnd. skewness = -0,442787
Stnd. kurtosis = -0,655697

Se observa que media y mediana son muy parecidas como ocurre con datos normales, La asimetría estandarizada (Std, Skewness) es un valor que está en el intervalo (-2,2) como pasa con datos simétricos. La curtosis estandarizada (Std.Kurtosis) está dentro del intervalo (-2,2) por lo que son datos normales.

Summary Statistics for TIEMPO

Count = 100
Average = 39,8454
Median = 26,7829
Standard deviation = 41,156
Stnd. skewness = 8,91972
Stnd. kurtosis = 12,4844

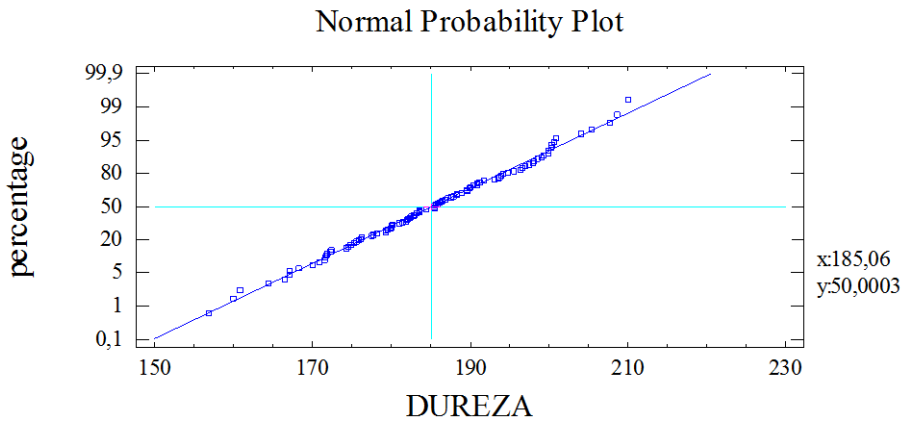
Sin embargo para la variable TIEMPO se observa que la media 39,84 es mayor que la mediana 26,78 como ocurre con datos con asimetría positiva. La asimetría estandarizada 8,91 está claramente por encima de 2, lo que confirma la asimetría positiva. La curtosis estandarizada 12,48 es mayor que 2 por lo que son datos leptocúrticos y no siguen distribución normal.

c) ¿Cómo podrías calcular aproximadamente sobre el Papel Probabilístico Normal la media y la desviación típica de las distribuciones DUREZA y TIEMPO?

Respuesta:

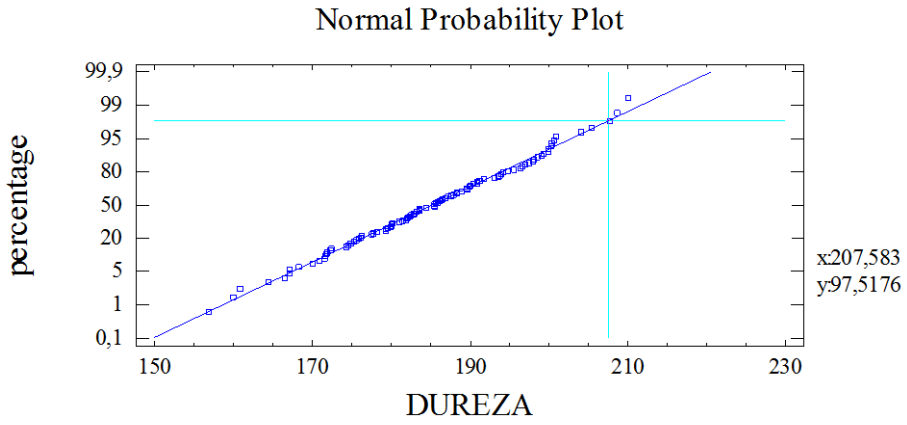
Se puede calcular la media y desviación típica de DUREZA sobre el papel probabilístico normal porque hay una recta que ajusta a los puntos. Para la variable TIEMPO no se puede calcular la media y la desviación típica con el papel probabilístico normal porque como se aprecia en el gráfico siguen una curva y no se puede ajustar una recta al no tener distribución normal.

Para calcular la media m y la desviación típica σ de DUREZA de forma aproximada se utilizarán las propiedades de la distribución normal. La media m en esa distribución coincide con la mediana, que es el percentil 50 es decir el valor que deja por debajo una probabilidad acumulada del 50%. Por tanto sobre el gráfico de papel probabilístico normal (Normal Probability Plot) hacer doble click con la techa izquierda del ratón para maximizar la ventana. Después con la tecla derecha del ratón abrir una ventana en la que estará activado **Locate**, seleccionarlo y a continuación mover las líneas auxiliares que aparecen de forma que la horizontal esté en Y:50% y mover la vertical hasta que corte la recta de los datos y la línea verticar. La coordenada X del punto de corte es aproximadamente la media. La figura siguiente muestra un posible resultado de esta operación:



Como se observa para Y:50% se lee X m≈185,06.

Para estimar aproximadamente la desviación típica σ a partir del papel probabilístico normal se puede utilizar la propiedad de la normal P(m-2σ<N(m,σ)<m+2σ)≈0,95. Así la probabilidad acumulada por debajo de m+2σ es aproximadamente 97,5%. Nos vamos al gráfico de papel probabilístico normal y con **Locate** colocamos la línea horizontal en Y cerca de 97,5 y la vertical que corte con ella y la línea de los datos. La coordenada X del punto de corte será aproximadamente m+2σ. Por ejemplo puede resultar:



Según el gráfico m+2σ≈207,583 como m≈185,06 entonces σ≈(207,583-185,06)/2=11,26

RECUERDA. La media (m) y la desviación típica (σ) obtenidas son las de la **población** (UD4). La media y desviación típica **muestrales** ($\bar{X}$ y S) se calcularían a partir de los datos (UD2).

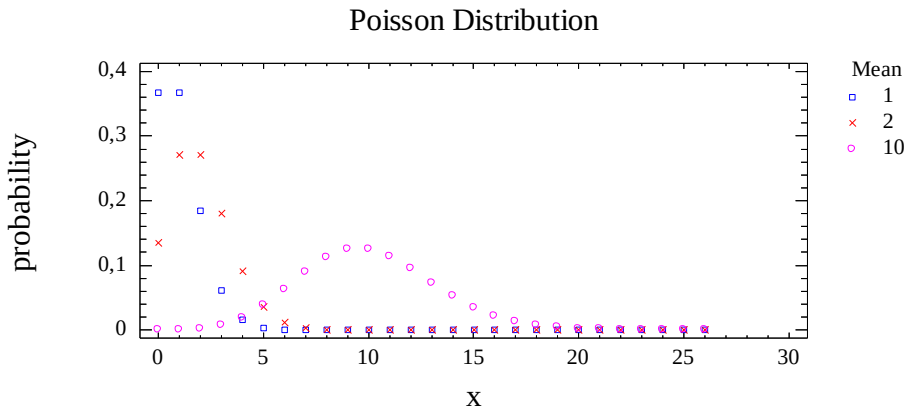
3. Aproximaciones normales

Construye la función de probabilidad P(x) de 3 variables de Poisson, una con λ=1, otra con λ=2 y una tercera con λ=10. ¿Qué se observa?

Para construir las tres funciones de probabilidad id a la opción **Plot…Probability Distributions** y en la lista de distribuciones que aparece seleccionad **Poisson**. Una vez abierta la ventana de análisis con la tecla derecha del ratón abrir la ventana en la que seleccionais **Analysis Options** y en ella poner en **Mean** los tres valores 1, 2 y 10 cada

uno en una casilla. Tras pulsar Ok aparecerá en el gráfico de la parte superior derecha las tres funciones de probabilidad.

Resultado:



Mean significa media y es el parámetro lambda λ de la distribución de Poisson. Se observa que cuando λ es mayor que 9 la función de probabilidad se aproxima a la distribución normal de campana de Gauss.

4. Teorema Central del Límite

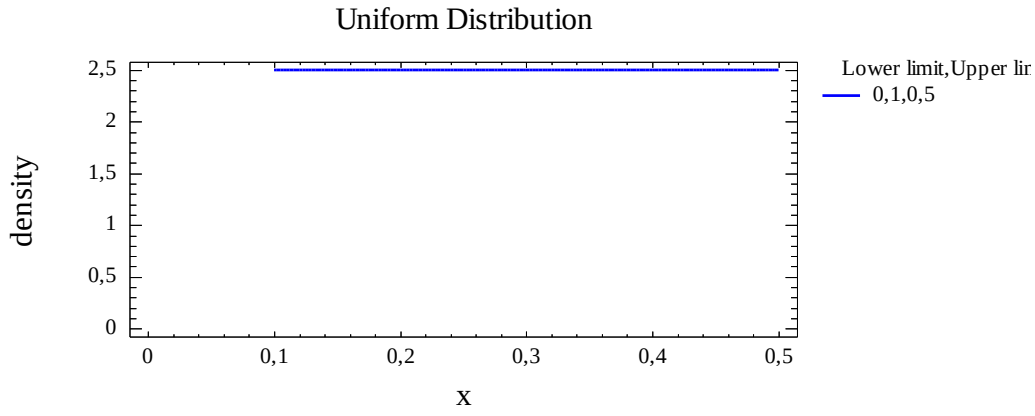
a) Observa la forma de la función de densidad f(x) que caracteriza a la distribución Uniforme de la variable aleatoria TIEMPO DE ACCESO A UN FICHERO del Ejercicio 11 de la UD4-Parte 3.

Opciones: *Plot → Probability Distributions (Uniform)*

Analysis Options: Lower limit=0,1 Upper Limit=0,5

Resultado:

En la ventana superior derecha aparece la función de densidad de la U(0,1, 0,5):



Se observa que es una línea recta horizontal y que encierra un área igual a un rectángulo.

b) ¿Cuál será el tiempo medio de acceso a un fichero? ¿Y la varianza?

RECUERDA. La media (m) y varianza (σ^2) obtenidas son las relativas a la población (UD4).

Resultado:

Para calcular la media y la varianza del tiempo de acceso a un fichero, como sigue distribución uniforme U(0,1,0,5), se utilizan las fórmulas de esta distribución.

Media de la uniforme $m=\frac{a+b}{2}=\frac{0,1+0,5}{2}=0,3$ s

Varianza de la uniforme $\sigma^2=\frac{(b-a)^2}{12}=\frac{(0,5-0,1)^2}{12}=0,013$ s^2

c) Obtener 10 muestras de la variable TIEMPO DE ACCESO A UN FICHERO de 100 datos cada una. Cada muestra puede obtenerse a partir de la generación de 100 datos aleatorios de una variable que fluctúa uniformemente entre 0,1 y 0,5 segundos.

Seleccionar ***Tabular Options: Random numbers***. Después seleccionar el icono **Save Results**. En la ventana que se abre seleccionar la variable a guardar y darle de nombre X1 (se pone el nombre en Target Variable). Hacer click en OK. Repetir este paso diez veces cambiando cada vez el nombre de la variable a X2, X3, X4, X5, X6, X7, X8, X9, X10. Se generaran así 10 muestras con 100 datos cada una de la U(0,1, 0,5) y se guardan el editor de datos. A cada uno se le generarán unos valores distintos porque son al azar y el programa utiliza un método que parte de una semilla de generación de números al azar que será distinta en cada ordenador.

d) Compara las medias ($\overline{X}_i$) y varianzas muestrales (S_i^2) de las 10 variables generadas con las obtenidas en al apartado **b)**

Opciones: *Describe → Numerical Data → Multiple-Variable Analysis…*

En **Data** poned las 10 variables generadas. Pulsad OK. Cuando se abra la ventana de análisis seleccionad **Tabular Options** y dentro de esta ventana **Summary Statistics**. Una vez abierta esta ventana pulsad la tecla derecha del ratón para acceder a **Pane Options** y de los parámetros que aparecen seleccionad la media **Average** y la varianza Variance.

Observad que las 10 medias muestrales están cercanas a la teórica de la distribució calculada en el apartado **b)** m=0,3. Las 10 varianzas muestrales estarán próximas a la teórica de la distribución calculada en el apartado **b)** σ^2=0,013.

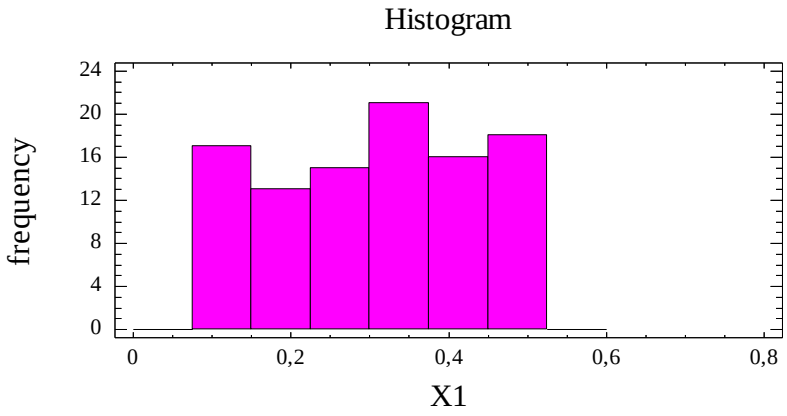
e) Construye un histograma con una de las variables uniformes generadas.

Con la opción **Plot…Exploratory Plots…Frequency Histogram**

En **Data** poned el nombre de una de las 10 variables generadas-

Resultado:

Por ejemplo a mi me ha dado un histograma para una de las variables con esta forma:



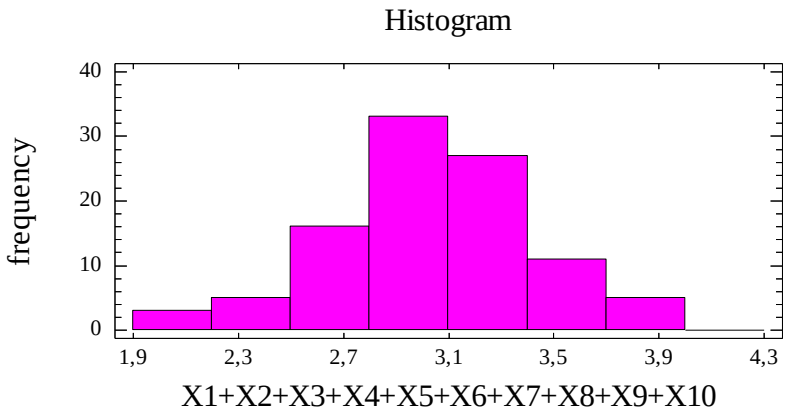
Se aprecia que no se parece en nada a la normal.Se parece más a un rectángulo que es la forma característica de los datos uniformes.

f) Genera una nueva variable X como suma de las 10 variables uniformes generadas, construye un histograma y compáralo con el obtenido en el apartado anterior. ¿Qué se observa?

Volved a entrar en la opción **Plot…Exploratory Plots..Frequency Histogram** pero ahora en **Data** poned X1+X2+X3+X4+X5+X6+X7+X8+X9+X10.

Resultado:

El histograma que me sale a mi para la suma con la generación que ha hecho mi ordenador de los números aleatorios es:



Para la suma de 10 uniformes independientes se aproxima a la distribución normal. El resultado teórico que justifica este resultado es el Teorema Central del Límite estudiado en clase de teoría.