Práctica de Estadística con Statgraphics

Teorema Central del Límite

Fundamentos teóricos

Importancia de la Distribución Normal

La distribución normal resulta especialmente importante porque aproxima lo observado en muchos procesos de medición. Por ejemplo, las medidas físicas del cuerpo humano en una población, las características psíquicas medidas por un test de inteligencia o personalidad, la nota de un examen de Bioestadística, las medidas de calidad en muchos procesos industriales, el tiempo de recuperación de una determinada lesión, o la mejoría experimentada en ciertos niveles biológicos después del tratamiento con un fármaco.

El Teorema Central del Límite

La justificación de la frecuente aparición de la distribución normal proviene del *Teorema Central del Límite*, que establece que cuando los resultados de un experimento son debidos a un conjunto muy grande de causas independientes que actúan sumando sus efectos, siendo cada efecto individual de poca importancia respecto al conjunto, es esperable que los resultados sigan una distribución normal.

En términos más matemáticos este teorema establece que si $X_1, ..., X_n$ son variables aleatorias independientes con distribución cualquiera (no necesariamente la misma), media μ_i y varianza σ_i^2, y construimos la variable suma:

$$Y = X_1 + ... + X_n$$

entonces, cuando n crece, la variable Y sigue una distribución normal cuya media es la suma de medias y cuya varianza es la suma de varianzas:

$$Y \sim N\left(\Sigma\mu_i, \sqrt{\Sigma\sigma_i^2}\right)$$

El valor de n variará según las variables aleatorias de las que se trate. Como valor orientativo suele darse $n \geq 30$. Sin embargo, si las variables de partida son continuas, simétricas y están idénticamente distribuidas, la aproximación se consigue para valores de n mucho más pequeños. Para distribuciones discretas puede ser necesario incluso un valor de n mayor o igual que 30 para tener una buena aproximación.

Aplicaciones del Teorema Central del Límite

Si aplicamos el Teorema Central del Límite a la suma de n variables independientes idénticamente distribuidas, y por lo tanto con igual media, μ, y desviación típica, σ, la variable suma resultante se distribuirá según una normal de media $n\mu$ y desviación típica $\sigma\sqrt{n}$:

$$Y \sim N\left(n\mu, \sigma\sqrt{n}\right)$$

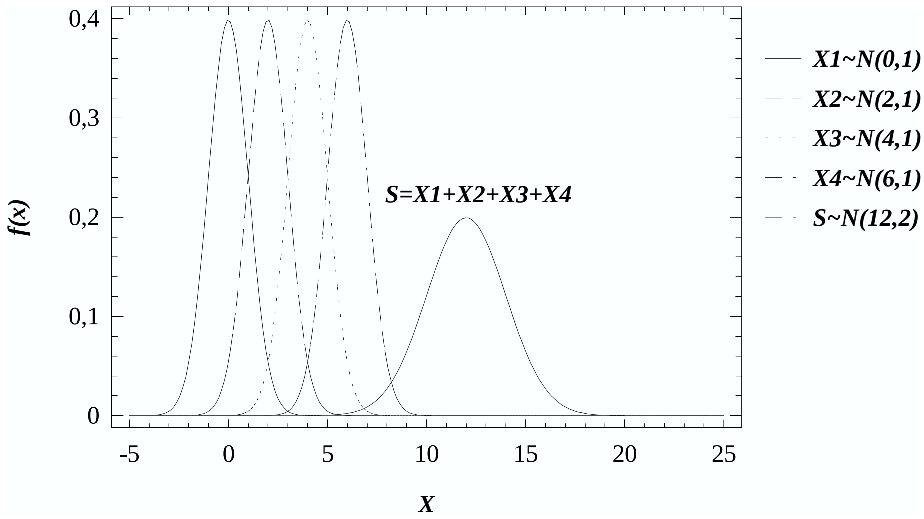

Figura 1: El Torema Central del Límite partiendo de 4 distribuciones normales.

Distribución para la Media Muestral

Si tenemos en cuenta que la *variable media muestral* resulta como suma de n variables idénticamente distribuidas:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

y aplicando propiedades de la media y la varianza, junto con el Teorema Central del Límite, puede demostrarse fácilmente que la media muestral sigue una distribución normal de media μ y desviación típica $\sigma/\sqrt{n}$:

$$\overline{X} \sim N\left(\frac{n\mu}{n}, \frac{\sigma\sqrt{n}}{n}\right) = N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Resultado este último muy importante y especialmente utilizado en Inferencia Estadística.

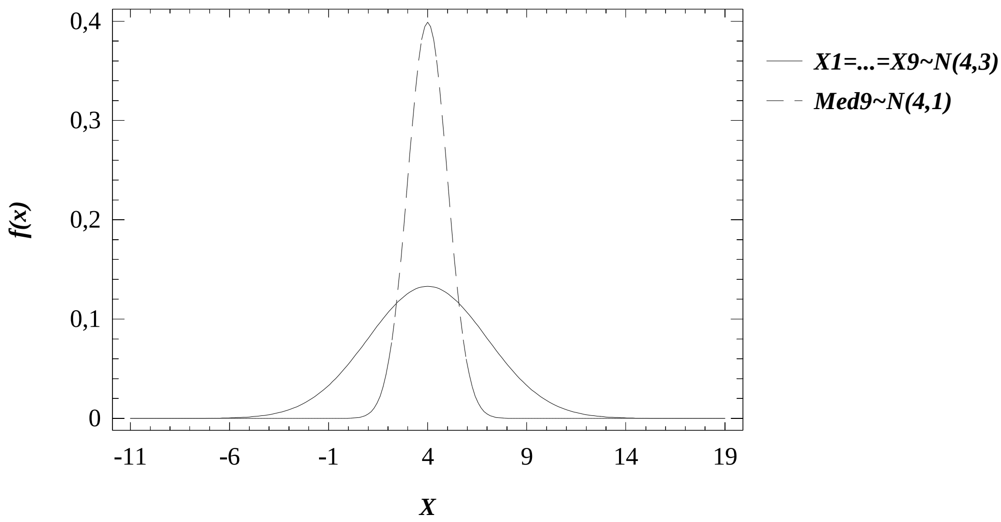

Figura 2: Distribución de la Media Muestral de 9 variables normales de media 4 y desviación típica 3.

Aproximación de la Binomial mediante una Normal

Si tenemos en cuenta que la variable binomial se obtiene como suma de n variables de Bernouilli, idénticamente distribuidas, cada una de ellas con valores 0 ó 1 (éxito o fracaso), de media p (probabilidad de éxito) y desviación típica $\sqrt{p(1-p)}$, y si suponemos que $n \geq 30$, $np > 5$ y $n(1-p) > 5$ (estas dos últimas condiciones para garantizar la simetría de la distribución), entonces podemos aplicar el Teorema Central del Límite y afirmar que la variable Binomial se podrá aproximar mediante una normal de media np y desviación típica $\sqrt{np(1-p)}$.

$$X \sim N\left(np, \sqrt{np\,(1-p)}\right)$$

El anterior resultado se conoce como *Teorema de De Moivre* y también se aplica muy frecuentemente en Inferencia Estadística.

Ejercicios Prácticos

1. Para comprobar el cumplimiento del Teorema Central de Límite partiendo de 4 variables con distribución uniforme entre $[0, 10]$, se pide:

   a) Crear la variable `Unif1` que contenga una muestra de 1000 valores generados aleatoriamente a partir de una variable aleatoria uniforme entre 0 y 10.

   Indicación
   1) Seleccionar el menú `Gráficos->Distribuciones de Probabilidad`.
   2) Seleccionar `Uniforme` en el cuadro de diálogo que aparece.
   3) Hacer click con el botón derecho del ratón sobre la ventana de resultados obtenida y seleccionar `Opciones de Análisis`.
   4) Introducir el valor del límite inferior de la Uniforme en el campo `Límite Inferior` y el del límite superior en el campo `Límite Superior`.
   5) Hacer click en el botón `Opciones Tabulares` y activar la casilla `Números Aleatorios`.
   6) Hacer click con el botón derecho del ratón sobre la ventana de resultados obtenidos y seleccionar `Opciones de Ventana`.
   7) Introducir el tamaño de la muestra en el campo `Tamaño`.
   8) Hacer click en el botón `Guardar Resultados`.
   9) En el cuadro de diálogo que aparece, activar la casilla `Números Aleatorios para Dist.1` y escribir el nombre de la variable a generar en el campo `Variable Destino`.

   b) Comprobar que los valores de la variable `Unif1` se distribuyen uniformemente entre 0 y 10. Para ello, dibujar el histograma de dicha variable haciendo 10 clases y comprobar que la frecuencia de cada clase es aproximadamente 100.

   Indicación
   1) Seleccionar el menú `Descripción->Datos Numéricos->Análisis Unidimensional`.
   2) Seleccionar la variable `Unif1` en el campo `Datos` del cuadro de diálogo.
   3) Hacer click en el botón `Opciones Gráficas` y activar la casilla `Histograma`.
   4) Hacer click con el botón derecho del ratón sobre el histograma obtenido y seleccionar `Opciones de Ventana`.
   5) Introducir el número de clases y los límites superior e inferior de la partición en los campos correspondientes.

   c) Calcular la media y la desviación típica de la variable Unif1 y comprobar si se parecen a media y desviación típica de una variable aleatoria uniforme [0,10]. (recordar que la media de una distribución uniforme [a,b] es $(b+a)/2$ y su desviación típica $(b-a)/\sqrt{12}$)).

   Indicación
   1) En la misma ventana del apartado anterior hacer click en el botón `Opciones Tabulares` y activar la casilla `Resumen Estadístico`.
   2) Hacer click con el botón derecho del ratón sobre los resultados obtenidos y seleccionar `Opciones de Ventana`.
   3) Activar las casillas correspondientes a los estadísticos que se piden.

d) Crear tres variables más Unif2, Unif3 y Unif4 que contengan otras 3 muestras de 1000 valores generados aleatoriamente a partir de una variable aleatoria uniforme entre 0 y 10.

e) Crear la variable Suma que contenga la suma de los valores de las muestras generadas previamente.

Indicación

1) Con la variable Suma seleccionada en el editor de datos, seleccionar el menú `Edición->Generar Datos`.
2) En el editor de fórmulas, escribir la expresión que da origen a la nueva variable.

f) Dibujar el histograma que la variable Suma. ¿Qué forma presenta el histograma? ¿A qué distribución se parece?

Indicación

1) Seleccionar el menú `Descripción->Datos Numéricos->Análisis Unidimensional`.
2) Seleccionar la variable Suma en el campo `Datos` del cuadro de diálogo.
3) Hacer click en el botón `Opciones Gráficas` y activar la casilla `Histograma`.

g) Calcular la media, la desviación típica, el coeficiente de asimetría y el coeficiente de curtosis de la variable Suma. Si se cumple el Teorema Central del Límite, ¿cuáles deberían ser los valores estos estadísticos? ¿Hay grandes diferencias entre los valores que predice dicho teorema y los realmente obtenidos?

Indicación

1) En la misma ventana del apartado anterior hacer click en el botón `Opciones Tabulares` y activar la casilla `Resumen Estadístico`.
2) Hacer click con el botón derecho del ratón sobre los resultados obtenidos y seleccionar `Opciones de Ventana`.
3) Activar las casillas correspondientes a los estadísticos que se piden.

h) Crear la variable Media que contenga la media de los valores de las muestras generadas previamente.

i) Dibujar el histograma que la variable Media. ¿Qué forma presenta el histograma? ¿A qué distribución se parece?

Indicación

1) Seleccionar el menú `Descripción->Datos Numéricos->Análisis Unidimensional`.
2) Seleccionar la variable Media en el campo `Datos` del cuadro de diálogo.
3) Hacer click en el botón `Opciones Gráficas` y activar la casilla `Histograma`.

j) Calcular la media, la desviación típica, el coeficiente de asimetría y el coeficiente de curtosis de la variable Media. Comparar la media y desviación típica obtenidas con las de las variables de partida. ¿Se cumple el Teorema Central del Límite aplicado a la distribución de la media muestral?

Indicación

1) En la misma ventana del apartado anterior hacer click en el botón `Opciones Tabulares` y activar la casilla `Resumen Estadístico`.
2) Hacer click con el botón derecho del ratón sobre los resultados obtenidos y seleccionar `Opciones de Ventana`.
3) Activar las casillas correspondientes a los estadísticos que se piden.

2. Para comprobar que la aproximación de la binomial mediante una normal se deduce del Teorema Central del Límite, repetir los apartados del ejercicio anterior generando las muestras a partir de una distribución binomial de parámetros $n = 100$ y $p = 0{,}4$.

Problemas

1. Crear 4 variables aleatorias que contengan muestras de 1000 valores generadas aleatoriamente a partir de distribuciones $N(10\,,5)$, $P(10)$, $B(20\,,0{,}5)$ y $U(5\,,15)$ respectivamente y repetir los pasos de los ejercicios anteriores para ver si se cumple el Teorema Central del Límite.