ESTADÍSTICA II

Práctica nº 2 – Teoría de Muestras

Finalidad de la práctica:

1. Determinación del tamaño muestral
2. Seleccionar muestras mediante muestreo aleatorio simple con reposición
3. Seleccionar muestras mediante muestreo aleatorio simple sin reposición
4. Simulación de distintas variables aleatorias
5. Estudio de la distribución de un estadístico: Método de Monte Carlo e ilustración del Teorema Central del Límite

1. Determinación del tamaño muestral

Una cuestión importante en la toma de muestras es la determinación de su tamaño. Para ello es necesario proporcionar el error máximo permitido (e) y el nivel de confianza (1-α). El primero indica cuál es la máxima diferencia admitida entre el estadístico y el parámetro que aproximamos; y el segundo representa la probabilidad con la que aseguramos el resultado sobre el error.

En esta práctica calcularemos el tamaño muestral con diferentes escenarios. Determinaremos el tamaño muestral de una m.a.s. con reposición cuando el parámetro que se desea estimar es la media poblacional supuesto que no se conoce la distribución de la variable de partida, supuesto que se puede aplicar el T.C.L y posteriormente, cuando el muestreo es sin reposición. Repetiremos los cálculos con las fórmulas apropiadas en el caso de que el parámetro que se desea estimar sea la proporción muestral.

2. Seleccionar muestras mediante muestreo aleatorio simple con reposición

En esta parte de la práctica vamos a seleccionar los elementos de una población de tamaño N que van a formar parte de una muestra aleatoria simple con reposición de tamaño n, para ello aplicaremos el siguiente procedimiento:

1. **Numerar los elementos** de la población de 1 a N.
2. **Generar un número aleatorio** con la función **ALEATORIO()** de Excel 2003

3. **Seleccionar** para la muestra la unidad de la población numerada con el valor surgido de la generación. Para ello y como los números generados provienen de una distribución U(0,1), los multiplicamos por N, les sumamos 1 y nos quedamos con la parte entera del número (En Excel 2003 calcularíamos **ENTERO(ALEATORIO()*N+1)**
4. **Repetir el proceso** n veces para conseguir una muestra del tamaño deseado.
5. **Conformar la muestra.** Una vez localizados todos los elementos de la población que deben formar parte de la muestra, se utiliza la función **BUSCARV(valor_buscado; matriz_buscar_en; indicador_columnas;ordenado)** para extraerlos de la misma.

3. Seleccionar muestras mediante muestreo aleatorio simple sin reposición

Si el tipo de muestreo que realizamos es un aleatorio simple sin reposición en poblaciones finitas, simularemos para cada individuo de la población una variable aleatoria Bernoulli cuyo parámetro coincide con la **fracción de muestreo** (f = n / N) (usando la función de Excel **BINOM.CRIT(n;p;probabilidad)**). Aquel elemento de la población cuyo valor sea 1 formará parte de la muestra y si el valor que le corresponde es 0 no se incluirá en la misma. Como es un método de simulación no podemos asegurar "exactamente" el tamaño muestral, por lo tanto, necesitaremos confirmar que el tamaño muestral real sea igual o mayor que el teórico con un pequeño margen para cumplir con las condiciones exigidas en el muestreo.

4. Simulación de distintas variables aleatorias

Como ya se vio en la práctica 1, el método más conocido y utilizado para generar valores artificiales de una v.a. es el método de la Función de Distribución Inversa. En esta práctica se simularán valores de diferentes variables aleatorias tales como una uniforme U(a,b), una normal N(μ, σ), una uniforme discreta U_D(N) o una binomial Bi(n, p).

5. Estudio de la distribución de un estadístico: Método de Monte Carlo e ilustración del Teorema Central del Límite

Con este apartado de la práctica se pretende ilustrar la aproximación de la distribución de un estadístico mediante simulación de muestras aleatorias de la población de partida y la construcción de histogramas del estadístico objeto de estudio.

Para ello cada celda representará un elemento de la muestra, por lo tanto, dado el tamaño muestral bastará con simular en una fila tantos valores de una variable aleatoria como dicho tamaño. Si necesitamos más de una muestra entonces deberemos repetir esta simulación en tantas filas como muestras se necesiten.

El método consiste en simular un número elevado de muestras, con el tamaño determinado y para cada muestra calcular el estadístico estudiado, disponiendo así de un número elevado de valores del estadístico. Utilizando la práctica nº 1 podemos construir el histograma de frecuencias o el polígono de frecuencias y aproximar la distribución de dicho estadístico. El cálculo de las características (media, varianza, asimetría y curtosis) nos pueden ayudar en esta aproximación.

En esta práctica aplicaremos este procedimiento para el cálculo de la distribución del estadístico **media muestral** por el método de Monte Carlo.

No obstante, y si el tamaño muestral es elevado sabemos además que la media muestral (u otros estadísticos que se construyen como suma de valores aleatorios) se puede aproximar a una distribución normal cuya media es la media poblacional y cuya varianza es inversamente proporcional al tamaño de la muestra:

Dada una m.a.s. con reposición:

$$\left(X_1, X_2, \ldots, X_n\right) \text{con } X_i \sim X \text{ con media } \mu = E[X], \sigma^2 = \mathrm{Var}(X)$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \Leftrightarrow Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

Por tanto, si utilizamos un tamaño muestral grande, podemos utilizar nuestro ejemplo de la media muestral para comparar su histograma de frecuencias con la función de densidad de la distribución normal, ilustrando de este modo el TCL como método para determinar la distribución de un estadístico.

Las siguientes funciones están disponibles en EXCEL:

- ALEATORIO( ) devuelve un valor al azar entre 0 y 1, es decir un valor de una variable uniforme(0,1).
- BINOM.CRIT(n;p;probabilidad) calcula el inverso de la función de distribución de una Binomial(n,p)
- BUSCARV(valor_buscado; matriz_buscar_en; indicador_columnas;ordenado) Busca un valor específico en la columna más a izquierda de una matriz y devuelve el valor en la misma fila de una columna especificada en la tabla.
- DISTR.NORM(x;media;dev_estándar;acum) calcula el valor de la función de densidad (si acum = 0) o de distribución (si acum =1) en el punto x para una distribución normal con parámetros μ= media y σ=dev_estándar.
- DISTR.NORM.INV(probabilidad;media;desv_estándar). Calcula el inverso de la función de distribución de una normal con una media y una desviación típica dadas.
- DISTR.NORM.ESTAND.INV(probabilidad). Calcula el inverso de la función de distribución de una normal estandar.
- FRECUENCIA(datos;grupos) Datos es una matriz de un conjunto de valores o una referencia a un conjunto de valores cuyas frecuencias desea contar. Si datos no contiene ningún valor, FRECUENCIA devuelve una matriz de ceros.