Aproximaciones acerca del Teorema Central del Límite.

F. Garcia Castro; M.L. Casado Fuente; M. Barrero Ripoll;
L. Sebastian Lorente; M. A. Castejón Solanas.
(U. Politécnica de Madrid)

Entre los "paquetes de utilidades" que existen en el mercado, hemos elegido el programa STATGRAPHICS. Resulta relativamente fácil de asimilar, aún con pocos conocimientos del sistema operativo, y además cada vez es mayor el número de alumnos que eligen realizar su proyecto Fin de Carrera sobre temas de Teledetección y usan para desarrollarlo este paquete.

Las prácticas se realizan en un aula dotada con una red NOVELL que consta de un servidor, 16 puestos con ordenadores XT y 2 impresoras, además de 3 ordenadores con disco duro de 20 Mb.

Comentamos brevemente las prácticas realizadas, deteniéndonos sólamente en las que merezcan una atención especial por su interés para nuestros alumnos.

Práctica 1. Se motiva a los alumnos en la utilización del Statgraphics como ayuda para la realización de un trabajo propuesto en clase: Recogida de datos reales, preferentemente medidas relacionadas con la Topografía, y tratamiento matemático de dichos datos.

Nosotros mismos, partimos de datos que ellos nos proporcionan obtenidos en otras asignaturas de la carrera, y , empleando estos ejemplos, damos un repaso a las distintas posibilidades que el Statgraphics ofrece para el análisis de dichos datos.

La simplicidad en el manejo del paquete, la rapidez de ejecución y la calidad de los gráficos que se obtienen atrapan inmediatamente el interés del alumno, que tiene luego la posibilidad de trabajar por su cuenta en el aula de informá-

tica para tal efecto.

Práctica 2. Consiste realmente en la realización del trabajo de matemáticas al que hacíamos referencia. Después de la introducción al Statgraphics que hacemos en la práctica 1, ahora los alumnos la emplean para obtener las medidas descriptivas de la muestra que estén estudiando.

Se les insiste en que analicen los datos desde las perspectivas numérica y gráfica e interpreten los resultados que se obtienen.

Práctica 3. La llamamos "Ley de errores"-Distribución Normal.

Si bien en general esta distribución es importante debido a que se presenta en un gran número de fenómenos aleatorios reales, lo es especialmente para nuestros alumnos. Tanto en su formación como después en su ejercicio profesional, una de sus tareas fundamentales es "medir".

En todas las mediciones se cometen errores y teniendo en cuenta que bajo ciertas hipótesis, la ley normal da la distribución de los errores aleatorios, se comprende la importancia del conocimiento y manejo de esta distribución.

En esta práctica los alumnos han recogido previamente los resultados obtenidos en el curso anterior en la realización de la práctica de la asignatura de Instrumentos Topográficos, consistente en la observación de medidas angulares en un triángulo y posterior cierre.

Una vez conseguidos los errores de cierre de los distintos grupos de prácticas, construyen la distribución de frecuencias, obteniendo la media muestral μ y la varianza muestral σ^2 comparando con la correspondiente $N(\mu, s)$ mediante la gráfica de la función de densidad.

Práctica 4. "Aproximación al Teorema Central del Límite"

La realidad de que muchas variables aleatorias se dis-

tribuyan normalmente, se fundamenta en el Teorema Central del Límite.

La ley normal aproxima bien sumas de gran número de variables aleatorias independientes, y ocurre frecuentemente en la naturaleza que un efecto aleatorio no despreciable, es en realidad producido por una multitud de causas que individualmente sólo producen efectos mínimos.

En la terminología actual, este teorema establece que, bajo condiciones adecuadas, **la distribución de la suma de gran número de variables aleatorias independientes, es aproximadamente normal.**

Entonces, cuando realizamos mediciones de una distancia, si consideramos como "valor más probable" la media de las observaciones, dichas medias están distribuidas normalmente, incluso cuando la distribución subyacente no sea normal, con la única limitación de que la varianza sea finita.

En esta práctica se recogen 100 observaciones de una misma distancia, y se observa la distribución de la media muestral. Una medida obvia de la precisión es la misma desviación típica $\sigma/10$.

Práctica 5. "Bondad de ajuste".

El planteamiento que hacemos es el siguiente:

Supongamos que disponemos de observaciones que proceden de n repeticiones independientes de un experimento, y que queremos determinar cómo se ajustan estos datos a un modelo probabilístico para el experimento.

Lo que hacemos es comparar los valores observados con los valores teóricos derivados del modelo.

Aun en el caso de que el modelo fuera totalmente correcto, se podría anticipar alguna discrepancia entre las frecuencias observadas y las esperadas debido a la variación aleatoria.

Se puede usar un contraste de bondad de ajuste para determinar si la discrepancia observada es demasiado grande para distribuirla al azar.

Un método para juzgar la concordancia entre los datos del modelo adoptado, consiste en construir una tabla de frecuencias observadas, f_j, y frecuencias esperadas según el modelo, e_j.

Consideramos como medida D de la distancia entre f_j y e_j a:

$$D = \sum_{j=1}^{k} \frac{(f_j - e_j)^2}{e_j}$$

La división por la frecuencia esperada e_j, compensa el hecho de que las desviaciones grandes; $(f_j - e_j)^2$, son más probables en las clases más amplias.

Sea d el valor observado en la distribución χ^2 de D.

El nivel de significación (NS) de los datos, es la probabilidad de obtener una discrepancia, D, tan grande como la observada, es decir,

$$NS = P(D \geq d)$$

Suponiendo que el modelo adoptado es correcto, y que las frecuencias esperadas, e_j, son todos grandes, entonces

$$NS = P(\chi^2_{(k-r-1)} \geq d)$$

Siendo d=valor del estadístico de bondad de ajuste.
k= nº de clases usadas para calcular **d**
r= nº de parámetros estimados a partir de los datos.

Las desviaciones grandes en las clases con frecuencias esperadas pequeñas afectan desfavorablemente a la precisión de la aproximación.

Teniendo en cuenta estas consideraciones, los alumnos realizan en esta práctica el test de ajuste para las prácticas 3 y 4.

Practica 6. "Regresión lineal"

Con frecuencia se realizan experiencias donde se miden dos características sobre cada elemento de una muestra aleatoria extraida de una población.

Así, a veces, una variable Y depende de la variable independiente X.

Consideramos nosotros el caso en el que exista una relación lineal $y = \alpha + \beta x + e$, donde e, el error de azar, suponemos que es una variable aleatoria de media cero.

Llamando a y b a los estimadores mínimo cuadráticos de α y β, la recta y=a+bx es la recta de regresión estimada de y sobre x siendo

$$b = \frac{S_{xy}}{S_x^2} \quad , \quad a = \bar{y} - b\bar{x}$$

Para especificar la distribución de los estimadores a y b, suponemos que los errores aleatorios son variables aleatorias $N(0,\sigma)$ independientes y que $Y_i \sim N(\alpha + \beta x_i, \sigma)$.
Establecemos entonces que :

$$b \sim N(\beta, \sigma/\sqrt{S_x^2})$$

$$a \sim N(\alpha, \sigma\sqrt{\frac{\Sigma x_i^2}{nS_x^2}})$$

Llamamos suma de los cuadrados de los residuos

$$SCR = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \quad , \text{ y entonces } \quad SCR/\sigma^2 \sim \chi^2_{n-2}$$

(Ji-cuadrado de Pearson con n-2 grados de libertad)

Se observa a partir de los resultados anteriores que:

$$\sqrt{\frac{(n-2)S_x^2}{SCR}}(b - \beta) \sim t_{n-2}$$

(t de Student con n-2 grados de libertad)

Así, podemos obtener un intervalo de confianza al $100(1-\lambda)\%$ para estimar β

Análogamente para α se obtiene:

$$a \pm \sqrt{\frac{\sum_{i=1}^{n} x_i^2 SCR}{n(n-2)\ S_x^2}}\ t_{\lambda/2,\,n-2}$$

y para $\alpha + \beta x_0$:

$$(a + bx_0) \pm \sqrt{\frac{1}{n} + \frac{(x_0 - x)^2}{S_x^2}}\ \sqrt{\frac{SCR}{(n-2)}}\ t_{\lambda/2,\,n-2}$$

de puntos, que les sugiere la recta de regresión y van efectuando los cálculos anteriores,

$$\bar{x} \ , \quad S_{xy} \ , S_y^2 \quad , \ S_x^2 \ , \ SCR.$$

haciendo las correspondientes interpretaciones acerca de la relación existente entre X e Y.

Por último calculan los intervalos de confianza.